

Original document

Speech intensifying-characteristic weighing-logrithmic spectrum addition method for anti-noise speech recognition

Publication number: CN1397929

Publication date: 2003-02-19

Inventor: CAO ZHIGANG (CN); XU TAO (CN)

Applicant: UNIV TSINGHUA (CN)

Classification:


- international: **G10L15/00; G10L15/20; G10L15/00; (IPC1-7):**
G10L15/00; G10L15/20

- European:

Application number: CN20021024144 20020712

Priority number(s): CN20021024144 20020712

Also published as:

 CN1162838C (C)

[View INPADOC patent family](#)

[View list of citing documents](#)

[Report a data error here](#)

Abstract of CN1397929

A "speech intensifying (MMSE)-feature weighting (FW)-logrithmic spectrum addition (LA)" method for anti-interference speech recognition features that according to the speech features in short time, the local S/N ratio is extracted, the confidence of the feature, that is weight, is estimated, and the recognition algorithm is such modified that the weight information is used. Its advantages are high S/N ratio and high recognition percentage up to 80% in strong noise condition.

Data supplied from the *esp@cenet* database - Worldwide

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.⁷
G10L 15/00
G10L 15/20



[12] 发明专利申请公开说明书

[21] 申请号 02124144.9

[43] 公开日 2003 年 2 月 19 日

[11] 公开号 CN 1397929A

[22] 申请日 2002.7.12 [21] 申请号 02124144.9

[71] 申请人 清华大学

地址 100084 北京市 100084 - 82 信箱

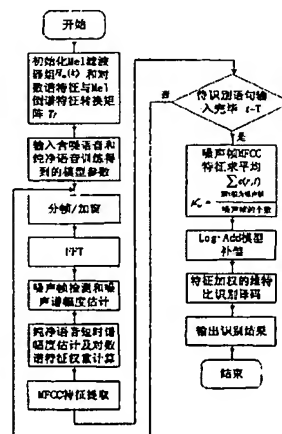
[72] 发明人 曹志刚 许 涛

权利要求书 5 页 说明书 21 页 附图 13 页

[54] 发明名称 抗噪声语音识别用语音增强 - 特征加权 - 对数谱相加方法

[57] 摘要

抗噪声语音识别用语音增强 - 特征加权 - 对数谱相加方法属于语音识别技术领域,其特征在于:它是一种融合多空间抗噪声语音识别技术,即 MMSE(最小均方差增强) - FW(特征加权) - LA(对数谱相加)的方法,它根据短时段语音各维特征提取空间的局部信噪比,给出特征的置信度估计,即权重,并对识别算法进行修改,把权重信息代入识别过程。尤其是前端语音增强技术、特征加权和对数谱相加模型补偿算法分别针对噪声在信号、特征和模型空间造成的失配进行处理,从而整体地提高了语音识别系统的抗噪声性能。在 SNR(信噪比)为 -5dB 的高斯白噪声和汽车噪声这种强背景噪声环境下,识别率都达到了 80%,而且前端增强和特征权重估计相互融合,选用了计算量较低的 MMSE 法,模型补偿也不需要噪声模型进行离线估计。



1. 抗噪声语音识别用语音增强—特征加权—对数谱相加方法, 含有计算机上运行的语音增强—对数谱相加方法, 其特征在于, 它依次含有以下步骤:

(1). 初始化 Mel 滤波器组在各线性频点 k 上的抽头系数 $H_m(k)$, 以及对数谱特征与 MFCC(Mel 频段倒谱系数)特征的转换矩阵 Tr 和 Tr^{-1} : 其中 $k=1,2,\dots,N_{fft}/2$, N_{fft} 是 FFT 的频点数; $m=1,2,\dots,M$, M 是 Mel 滤波器的个数。

(2). 输入含噪语音和纯净语音经训练得到的模型参数:

μ^c : 纯净语音训练得到的模型状态在 MFCC 倒谱域下的静态特征均值;

$\Delta\mu^c$: 纯净语音训练得到的模型状态在 MFCC 倒谱域下的动态特征均值;

(3). 分帧、加窗:

若采样后的原始语音为 $y(n)$, 汉明(hamming)窗在第 n 个采样点上的系数:

$$h(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), n=1, \dots, N$$

N 等于帧长, 则分帧后的原始语音信号为:

$$y(n,t) = y\left(\frac{N \times (t-1)}{2} + n\right), n=1, \dots, N$$

t 表示帧号, 加上汉明窗后的原始语音信号为:

$$y_w(n,t) = y(n,t) \times h(n), n=1, \dots, N$$

(4). 快速傅立叶变换 FFT:

由于语音短时频谱对感知语音起决定性的作用, 利用 FFT 逐帧将语音变换到频谱域:

$$\bar{Y}(k,t) = Y(k,t)e^{j\angle Y(k,t)} = FFT\{y_w(n,t)\}, k=1, \dots, N_{fft}$$

N_{fft} 是 FFT 变换的点数。

(5). 噪声帧检测和噪声谱幅度估计:

(5.1). 设定前 10 帧起始段含噪语音为噪声帧, 输入当前第 t 帧含噪语音的短时谱幅度:

(5.2). 若当前帧为起始段噪声帧, 则前 t 帧噪声功率谱幅度的估计值为:

$$\bar{D}_p(k,t) = \left[\sum_{s=1}^t Y(k,s)/t \right]^2$$

并在当前帧为第 10 帧时输出起始段噪声谱幅度的估计值:

$$N(k) = \sum_{s=1}^{10} Y(k,s)/10$$

计算用于区分噪声帧和含噪语音帧的判决门限 χ :

$$\chi = \text{Max}_{t=1,2,\dots,10} \left\{ \sum_{k=1}^{N_p/2+1} \text{Pow}[Y(k,t)/N(k),5] \right\}$$

(5.3).若当前帧不是起始段噪声帧,则当前帧 t 的判决值:

$$\rho = \sum_{k=1}^{N_p/2+1} \text{Pow}[Y(k,t)/N(k),5]$$

(5.3.1)若 $\rho < \chi$,则判决为含噪语音中的噪声帧,其噪声功率谱幅度估计值为:

$$\tilde{D}_p(k,t) = 0.98 \times \tilde{D}_p(k,t-1) + 0.02 \times Y_p(k,t)$$

并输出:

(5.3.2).若 $\rho \geq \chi$,则判决为非噪声帧,即含有噪声的语音帧,其噪声功率谱幅度为:

$$\tilde{D}_p(k,t) = \tilde{D}_p(k,t-1)$$

并输出:

(6).用取决于先验信噪比 ζ 和后验信噪比 γ 的谱幅度增益系数 $G(k,t)$ 计算纯净语音短时谱幅度的估计值,以及相应的第 t 帧第 m 个对数谱特征的权重 $w_m(t)$:

(6.1).输入当前第 t 帧含噪语音的短时谱幅度:

(6.2).计算当前帧 t 第 k 个频点的后验信噪比 $\gamma(k,t) = Y_p(k,t)/\tilde{D}_p(k,t)$, $Y_p(k,t)$ 为含噪语音的功率谱幅度, $\tilde{D}_p(k,t)$ 为估计的噪声功率谱幅度。

(6.2.1).如果当前帧 $t=1$,则初始化当前帧第 k 个频点的先验信噪比为 $\zeta(k,t)=0.1$;

(6.2.2).如果当前帧 $t>1$,则利用上一帧的先验和当前帧的后验信噪比,通过滑动平均估计得到当前帧第 k 个频点的先验信噪比:

$$\zeta(k,t) = 0.98 \times \zeta(k,t-1) + 0.02 \times [\gamma(k,t)-1]$$

(6.3).当前帧 t 第 k 个频点的谱幅度增益系数:

$$G(k,t) = \frac{1}{2} \sqrt{\frac{\pi \zeta(k,t)}{\gamma(k,t)(1+\zeta(k,t))}} \Psi(-0.5; 1; -\frac{\gamma(k,t)\zeta(k,t)}{1+\zeta(k,t)})$$

利用级数求和,计算得到:

$$\Psi(a_1, a_2, a_3) = 1 + \frac{a_1}{a_2} \frac{a_3}{1} + \frac{a_1(a_1+1)}{a_2(a_2+1)} \frac{a_3^2}{2!} + \dots$$

其中 $a_1 = -0.5$, $a_2 = 1$, $a_3 = -\frac{\gamma(k,t)\zeta(k,t)}{1+\zeta(k,t)}$

(6.4).相应的纯净语音短时谱幅度的估计值为:

$$\hat{X}(k,t) = G(k,t)Y(k,t)$$

(6.5).重新计算当前帧第 k 个频点的先验信噪比:

$$\zeta(k, t) = |\hat{X}(k, t)|^2 / \tilde{D}_p(k, t)$$

(6.6).计算完当前帧 t 第 k 个频点($1 \leq k \leq N_{\text{fft}}/2+1$)的 $G(k, t)$ 、 $\hat{X}(k, t)$ 和 $\zeta(k, t)$ 值。

(6.7).计算当前帧 t 第 m 个对数谱特征的权重:

$$w_m(t) = \sum_{k=1}^{N_{\text{fft}}/2} G(k, t) H_m(k) / \sum_{k=1}^{N_{\text{fft}}/2} H_m(k)$$

(6.8).计算当前帧共 M 个对数谱特征的的权重, M 是对数谱特征的维数。

(6.9).计算完 $t=1, 2, \dots, T$ 各帧中的 $\hat{X}(k, t)$ 和 $w_m(t)$;

(6.10).输出所有相应的纯净语音短时谱幅度估计值 $\hat{X}(k, t)$ 和对数谱特征的权重

$w_m(t)$;

(7).MFCC 特征提取

(7.1).输入纯净语音短时谱幅度估计值 $\hat{X}(k, t)$;

(7.2).计算功率谱: $\hat{X}_p(k, t) = |\hat{X}(k, t)|^2, k=1, \dots, N_{\text{fft}}$;

(7.3).Mel 滤波:

$$M\text{Bank}(m, t) = \sum_{k=1}^{N_{\text{fft}}/2} H_m(k) \times \hat{X}_p(k, t), m=1, \dots, M$$

(7.4).对数谱特征: $F\text{Bank}(m, t) = \log(M\text{Bank}(m, t)), m=1, \dots, M$

(7.5).DCT 倒谱表示:

$$\tilde{c}(r, t) = \alpha(r) \sum_{m=1}^M F\text{Bank}(m, t) \cos\left(\frac{\pi(2m-1)(r-1)}{2M}\right), r=1, \dots, M$$

其中 $\alpha(1) = \sqrt{\frac{1}{M}}, \alpha(r) = \sqrt{\frac{2}{M}}, r=2, \dots, M$, 并取前 R 维特征矢量

(7.6).倒谱加权:

$$c(r, t) = \text{lifter}(r) \times \tilde{c}(r, t), r=1, \dots, R$$

其中 $\text{lifter}(r) = 1 + \frac{L}{2} \sin\left(\frac{\pi(r-1)}{L}\right), r=1, \dots, R$, L 为加权滤波器宽度;

(7.7).计算动态系数:

$$\Delta c(r, t) = \sum_{\Delta t=-2}^2 \Delta t c(r, t + \Delta t) / 10, \Delta t \text{ 表示帧间距};$$

(7.8).输出 $c(r, t)$ 和 $\Delta c(r, t)$;

(8).判断待识别语句是否输入完毕 $t=T$?

(9).若判断为待识别语句已经输入完毕, 则计算噪声帧, 即剩余噪声的静态 MFCC 特征平均值, 剩余噪声的定义如下:

$$\hat{d}(n) = \hat{x}(n) - x(n)$$

其中 $x(n)$ 表示纯净语音在第 n 个样点上的值, $\hat{x}(n)$ 表示 $x(n)$ 增强后的估计值。由于剩余噪声存在于各个语音帧, 而语音仅存在于非噪声帧, 所以对于噪声帧来说, $\hat{D}(k, t) = \hat{X}(k, t)$, 即剩余噪声的短时谱幅度在各噪声帧中等于增强后语音的短时谱幅度, 我们可以利用下式计算剩余噪声的静态 MFCC 特征均值:

$$\mu_{nr}^c = \frac{\sum_{\text{第 } r \text{ 帧为噪声帧}} c(r, t)}{\text{噪声帧的个数}}$$

其中噪声帧包括起始段 10 帧和后面判决的噪声帧, $r = 1, 2, \dots, R$ 。

(10).Log-Add 对数谱相加模型补偿:

(10.1).输入剩余噪声的 MFCC 特征均值并转换到对数谱域 $\mu_n^l = Tr^{-1} \mu_n^c$;

(10.2).输入纯净语音训练模型的状态均值, 并转换到对数谱域 $\mu^l = Tr^{-1} \mu^c$, $\Delta \mu^l = Tr^{-1} \Delta \mu^c$;

(10.3).Log-Add 模型补偿:

$$\hat{\mu}_m^l = \mu_m^l + \log(1 + \exp(\mu_{nm}^l - \mu_m^l)), m = 1, 2, \dots, M$$

$$\Delta \hat{\mu}_m^l = \frac{\Delta \mu_m^l}{1 + \exp(\mu_{nm}^l - \mu_m^l)}$$

(10.4).把补偿的模型状态转换到 MFCC 倒谱域 $\hat{\mu}^c = Tr \hat{\mu}^l$, $\Delta \hat{\mu}^c = Tr \Delta \hat{\mu}^l$;

(10.5).当状态输入完毕, 输出剩余噪声补偿后的语音模型;

(11).特征加权的维特比识别译码:

(11.1).输入剩余噪声补偿后的语音模型、增强语音当前帧 MFCC 特征 y_i^c 、对数谱特征权重 $w_m(t)$;

(11.2).计算观测帧在候选状态下的对数概率似然值:

(11.2.1).在 MFCC 倒谱域计算 MFCC 特征与可选状态的状态均值的矢量差:

$$d^c = y_i^c - u^c;$$

(11.2.2).把差矢量变换到对数谱特征域: $d^l = Tr^{-1} d^c$;

(11.2.3).在对数谱域进行加权, 并变换回 MFCC 倒谱域 $\bar{d}^c = Tr W d^l$;

(11.2.4).计算对数概率似然值:

$$\log(p(y_i^c | q(t) = i)) = C(\Sigma^c) - \frac{1}{2} \bar{d}^{cT} (\Sigma^c)^{-1} \bar{d}^c$$

其中 Σ^c 为倒谱域的状态方差矩阵, 且为对角阵 $\Sigma^c = \text{Diag}\{\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{iR}\}$, c 表示倒谱

域, i 表示状态; $C(\Sigma^c)$ 表示与 y_i^c 无关的常数项, 对应 $-\sum_{r=1}^R \log(\sqrt{2\pi} \sigma_{ir})$, R 是倒谱

特征的维数。

(11.3).初始化 Viterbi 译码后,再迭代,计算完 $t=1,2,\dots,T$ 帧;

(11.4).计算最大概率 $p^* = \max_{1 \leq i \leq N} [\delta_T(i)]$ 和最佳路径的终止状态: $\hat{q}(T) = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$;

(11.5)通过回溯依次输出最佳路径上的其他状态: $\hat{q}(t) = \varphi_{t+1}(\hat{q}(t+1)), t = T-1, \dots, 1$;

(12). 输出识别结果, 结束。

抗噪声语音识别用语音增强—特征加权—对数谱相加方法

技术领域

抗噪声语音识别用语音增强—特征加权—对数谱相加方法属于语音识别技术领域。

背景技术

基于 HMM(Hidden Markov Model)的概率统计识别方法是目前自动语音识别(ASR: Automatic Speech Recognition)研究中最常用的模型框架。具有里程碑意义的 HMM 被引入语音识别领域, 由于它能较好的描述语音的产生机理, 并且有比较简明的模型估计(训练)与状态搜索算法, 极大的促进了语音识别技术的发展。

隐含马尔可夫模型可以看成是一个有限状态自动机, 见图 1, 这是一个最常用的 HMM 的拓扑结构。在每一个离散时刻, 对应任意第 t 帧语音, 它只能处于有限多种状态中的某一种状态。假设允许出现的状态有 U 种, 记之为 $S_u, u=1 \sim U$ 。若自动机在第 t 帧语音时所处的状态用 $q(t)$ 表示, 那么 $q(t)$ 只能等于 $S_1 \sim S_U$ 中的某一个, 这可表述为 $q(t) \in \{S_1 \sim S_U\}, \forall t$ 。如果此自动机在 $t=1$ 时开始运行, 那么以后每一帧所处的状态以概率方式取决于初始状态概率矢量 π 和状态转移概率矩阵 A 。对于任意帧 $t, (t \geq 1)$, 自动机的状态 $q(t)$ 取 $S_1 \sim S_U$ 中哪一种的概率只取决于前一帧 $t-1$ 时所处的状态, 而与更前的任意帧所取的状态无关。这样, 由此产生的状态序列 $q(1), q(2), q(3), \dots$ 是一条一阶马尔可夫链。此系统在任意帧 t 时所处的状态 $q(t)$ 隐藏在系统内部, 不为外界所见, 外界只能得到系统在该状态下提供的随机输出(在这里是语音信号), 隐含马尔可夫模型由此得名。

我们知道, 语音信号具有短时平稳特性。为此, 可以将语音划分为不同的短时段, 每段对应于 HMM 的一个状态, 段与段之间的迁移可以用 HMM 中状态到状态的转移来表示。每个状态具有特定的模型参数, 可以描述一帧语音的平稳的统计特性, 如果下一帧语音具有相同的统计特性, 则状态不转移, 或者说下一个状态仍然跳到本状态, 反之如果下一帧语音的统计特性变化了, 则下一个状态会跳到与该段语音统计特性相符的状态。

由上可以看出, 隐含马尔可夫模型是建立在一定物理意义上的数学模型, 其中的各状态相对于发音器官在人说话中所经历的每个相对稳定的过程, 比较贴切的描述了语音信号的时变性和准平稳性。图 1 示出了 HMM 对输入语音的描述。图中语音为中文的“他去无锡市”。我们同时用相应的音子来标注输入语音。各音子标注相对于一个 HMM。我们在图中示出了一个从左到右的 HMM 拓扑结构。各状态有相应的输出概率分布。状态 1 和状态 9 分别为起始状态和终止状态, 它们用来将不同的 HMM 串接起来, 只是一个不占时间的过渡状态, 本身并不产生对外的输出。我们用实线画出了不同标注划分的语音倒谱均值。

为表述方便, 直接用状态编号 i, j 表示状态集合 $\{S_1 \sim S_U\}$ 中的第 i 个和第 j 个状态, U 表

示模型状态总数。

A —状态转移概率矩阵，元素为：

$$a_{ij} = P(j|i), 1 \leq i, j \leq U \quad (1)$$

表示由状态 i 到状态 j 的概率。根据转移概率的定义，我们有，

$$\sum_{j=1}^U a_{ij} = 1, \forall 1 \leq i \leq U \quad (2)$$

在图 1 的最常用的具有由左到右拓扑结构的 HMM 中， A 实际上为一双线对角阵。

B —输出概率密度：

$$\begin{aligned} p(y_i | q(t) = i) &= N(y_i; \mu_i, \Sigma_i) \\ &= \prod_{r=1}^R \frac{1}{\sqrt{2\pi}\sigma_{ir}} \exp\left(-\frac{(y_{ir} - \mu_{ir})^2}{\sigma_{ir}^2}\right) \end{aligned} \quad (3)$$

表示在状态 $q(t) = i$ ，对于观测语音特征 y_i 的似然值。语音信号特征的概率分布可以用高斯函数来逼近，其中 $y_i = [y_{i1}, y_{i2}, \dots, y_{iR}]$ 是 R 维观测特征矢量， $\mu_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{iR}]$ ， $\Sigma_i = \text{diag}[\sigma_{i1}^2, \sigma_{i2}^2, \dots, \sigma_{iR}^2]$ 分别是高斯函数 $N(y_i; \mu_i, \Sigma_i)$ 的均值和方差，由于 $y_i = [y_{i1}, y_{i2}, \dots, y_{iR}]$ 一般是经过正交变换得到的，所以高斯分布的协方差矩阵用对角阵来描述，并且多维高斯分布可以写成多个一维高斯分布连乘的形式。

π —各状态的起始概率分布：

元素 $\pi_i \in [0, 1]$ 。在图 1 所示的 HMM 中，状态 1 是唯一的起始状态，所以 $\pi_1 = 1$ ，其余状态的起始概率均为 0。

以上参数是通过训练过程得到的。训练将通过训练语音数据来调整上述参数，也就获得了语音特征的统计信息。训练结束后，就可以进行识别了。

基于 HMM 的语音识别是将输入的语音特征序列 $Y = [y_1, y_2, \dots, y_T]$ ，根据最大似然准则，搜索出最佳状态序列 $\hat{Q} = [\hat{q}(1), \hat{q}(2), \dots, \hat{q}(T)]$ ，从而揭开 HMM 的隐含部分，其中 T 是待识别的语音的长度，即有 T 个语音帧的特征。这个问题的解决通常采用 Viterbi 算法。定义：

$$\delta_i(i) = \max_{q(1)q(2)\dots q(t-1)} \log[p(q(1)q(2)\dots q(t-1), q(t) = i, y_1 y_2 \dots y_t | \lambda)] \quad (4)$$

为给定模型参数，部分观测 $y_1 y_2 \dots y_t$ ，部分路径 $q_1 q_2 \dots q_{t-1}, q_t = i$ 的最大输出对数似然值，其中 λ 表示训练得到的 HMM 语音模型。

$$\text{初始化: } \delta_1(i) = \log \pi_i + \log[p(y_1 | q(1) = i)] \text{ 且 } \varphi_1(i) = 0 \quad (5)$$

$$\text{迭代: } \delta_i(j) = \max_{1 \leq i \leq N} [\delta_{i-1}(i) + \log(a_{ij})] + \log(p(y_t | q(t) = j)) \quad (6)$$

音特征，如感知线性预测系数(PLP: Perceptive Linear Predictive)。

三.模型空间处理。即利用噪声的统计特性，对理想环境下训练得到的语音模型进行校正，使之适用于特定的识别环境，如并行模型补偿(PMC: Parallel Model Compensation)和对数谱相加法(LA: Log-Add)。

这些方法在弱背景噪声环境下有效地提高了系统的识别性能，而在强背景噪声环境下识别精度还是急剧下降。本发明正是要解决低信噪比噪声环境下的语音识别问题。

把信号空间的最小均方误差(MMSE: Minimum Mean Square Error)增强处理和模型空间的对数谱相加(LA: Log-Add)补偿算法相融合，我们得到了一种解决方案，称之为 MMSE-LA 方案，它可以显著的提高低信噪比环境下的识别精度。本发明还在特征空间提出了一种新的特征加权算法，并利用 MMSE 增强方法，给出了有效的权重计算公式，从而提出了多空间信号处理的 MMSE-FW-LA 方案，FW 指特征加权(Feature Weight)，即同时在信号空间、特征空间和模型空间消除噪声引起的训练和识别环境的失配。

由于 MMSE-LA 和 MMSE-FW-LA 两种方案都涉及到 Mel 频段倒谱系数(MFCC: Mel Frequency Cepstral Coefficient)这一目前比较常用的声学特征，有必要事先予以介绍。

自动语音识别(ASR: Automatic Speech Recognition)是给定一段语音信号，由机器从中提取信息并确定语言含义的过程，它首先要从语音信号中提取能够反映语音本质、有利于识别并适于计算机处理的声学特征矢量。声学特征的发展经历了从时域到频域，再到倒谱域的过程，并且越来越多的结合了人耳听觉系统的知识。Mel 频段倒谱系数(MFCC: Mel-Frequency Cepstral Coefficient)是目前比较常用的声学特征。我们首先描述它的提取过程，如图 4 所示。

分帧和加窗 分帧利用了语音信号的短时平稳特性。通过分帧，可以把语音当作平稳随机信号进行分析。相邻的语音帧通过一定的重叠来保证各帧之间的相关信息。加窗的目的是减小频率混叠，通常是 Hamming 窗。

$$h(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), n = 1, \dots, N \quad (11)$$

其中 N 等于帧长， $h(n)$ 表示 hamming 窗在第 n 个样点上的系数。 $y(n)$ 表示采样后的原始语音，分帧后表示为：

$$y(n, t) = y\left(\frac{N \times (t-1)}{2} + n\right), n = 1, \dots, N \quad (12)$$

其中 t 表示帧号， n 表示当前帧的样点序号。加汉明窗之后表示为：

$$y_w(n, t) = y(n, t) \times h(n), n = 1, \dots, N \quad (13)$$

FFT 快速傅立叶变换 由于语音短时频谱对感知语音起决定性的作用，利用 FFT 逐帧将语音变换到频谱域，表达形式为：

$$\bar{Y}(k, t) = Y(k, t) e^{j\varphi(k, t)} = FFT\{y_w(n, t)\}, k = 1, \dots, N_{fft} \quad (14)$$

$$\varphi_i(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) + \log(a_{ij})] \quad (7)$$

$$\text{终止: 最大概率 } p^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (8)$$

$$\text{最佳路径的最后的的状态 } \hat{q}(T) = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \quad (9)$$

通过回溯依次求最佳路径上的其它路径:

$$\hat{q}(t) = \varphi_{t+1}(\hat{q}(t+1)), t = T-1, T-2, \dots, 1 \quad (10)$$

可以看出, $\delta_t(i)$ 用来记录在时刻 t 各状态产生部分输出的最大概率, 而 $\varphi_t(j)$ 则用来记录路径的连接信息。

目前纯净语音识别已达到一个比较成熟的阶段, 以 IBM 的 Via Voice 为代表, 对连续语音的识别率可达到 90% 以上, 但是对背景噪声和输入话筒有较严格的要求, 否则系统性能将会有很大的下降。造成这种情况的原因是训练环境和识别环境的失配。现在很多识别系统的参数都是在实验室环境中训练得到的, 训练语音大多是在安静背景下, 通过高质量麦克风采集的。而到了实际的应用场合, 由于多种因素的影响, 识别语音不可避免的会和系统参数存在失配, 从而造成实际性能和实验室中的性能的大相径庭。

造成语音识别中测试与训练环境的失配的原因有很多, 包括说话人本身的心情, 说话人周围的噪声, 录音时的信道, 录音时的背景噪声, 信号传递时的信道和接收端的背景噪声等。抗噪声语音识别只考虑接收背景噪声和卷积信道对语音信号的影响, 失配模型如图 2 所示。

目前抗噪声问题是语音识别领域中的一个热点。无处不在的噪声带来了训练环境和识别环境的失配, 从而造成识别器性能的急剧下降。抗噪声语音识别的目标就是要消除这种失配, 使识别性能尽可能的接近在训练环境下的性能。由于现在的语音识别系统普遍采用基于 HMM 的统计模型, 所以噪声带来的失配可以映射到如图 3 所示的三个空间。

在图 3 中, 训练和识别的失配表现在信号、特征值、模型三个空间。在信号空间, S_x 代表训练环境下的原始语音, S_y 代表识别环境下的语音, 两种环境下语音信号的失配由失真函数 $D_s()$ 表示。语音信号在经过特征提取过程后, 信号空间的失配必然也会表现到特征空间, F_x 是训练语音的特征, F_y 是测试语音的特征, 其失配用失真函数 $D_f()$ 来表示。最后, 特征 F_x 用来训练 HMM 得到模型 M_x , 而和特征 F_y 相匹配的模型应为 M_y , 这种在模型上的失配用失真函数 $D_m()$ 表示。

抗噪声语音识别的方法可以从图 3 中三个不同的角度来考虑, 在研究过程中, 基本形成了如下几类做法:

一. 信号空间的处理。采用信号处理方法提高语音识别系统抗噪声性能, 如利用语音增强技术和麦克风阵列来提高输入信号的信噪比。

二. 特征空间的处理。主要是结合人耳听觉的知识, 提取对噪声干扰不敏感的稳健性语

我们对 Mel-Scaled 滤波器组的输出取对数，得到对数谱特征参数(log-Spectra)。

$$FBank(m, t) = \log(MBank(m, t)), m = 1, \dots, M \quad (22)$$

其中 $FBank(m, t)$ 表示提取的第 t 帧语音的第 m 维对数谱特征。

离散余弦变换(DCT) DCT 具有类似正交变换的效果，能够使语音特征向量各维之间相关性减小；此外还能够使特征向量维数降低，进一步起到特征提取和特征压缩的作用。由于离散余弦变换使特征向量各维之间互不相关，所以可用对角阵来表示各维向量之间的协方差矩阵。在这种情况下，对角化的协方差矩阵对于计算来说相当于降低了一维，计算量大大降低，许多高效的算法可以得以实现。离散余弦变换定义为：

$$\tilde{c}(r, t) = \alpha(r) \sum_{m=1}^M FBank(m, t) \cos\left(\frac{\pi(2m-1)(r-1)}{2M}\right), r = 1, \dots, M \quad (23)$$

$$\alpha(1) = \sqrt{\frac{1}{M}}, \alpha(r) = \sqrt{\frac{2}{M}}, r = 2, \dots, M \quad (24)$$

其中， $\tilde{c}(r, t)$ 表示提取的第 t 帧语音的第 r 维倒谱系数。由于经过 DCT 变换后， M 维倒谱系数的后几维很小，因此可以降低特征向量维数，在识别计算中只取倒谱系数的前 R 维。

倒谱加权由于低维和高维的倒谱系数对噪声比较敏感，所以通常采用升余弦形式的带通函数对倒谱系数进行加权，在一定程度可以提高系统的稳健性。

$$lifter(r) = 1 + \frac{L}{2} \sin\left(\frac{\pi(r-1)}{L}\right), r = 1, \dots, R \quad (25)$$

其中 L 为加权滤波器宽度。加权后的倒谱系数为：

$$c(r, t) = lifter(r) \times \tilde{c}(r, t), r = 1, \dots, R \quad (26)$$

此加权过程称为倒谱滤波。 $c(r, t)$ 称为静态 MFCC 特征。

动态系数反映了语音谱中的动态信息。它们分别通过如下的公式计算而得：

$$\Delta c(r, t) = \sum_{\Delta t=-2}^2 \Delta t c(r, t + \Delta t) / 10 \quad (27)$$

其中 $\Delta c(r, t)$ 表示一阶 MFCC 特征系数， Δt 表示帧间距。

发明内容

本发明的目的在于提供一种低信噪比环境下使用的抗噪声识别用语音增强—特征加权—对数谱相加方法。

特征加权算法的出发点是认为噪声在不同时段，不同频率对语音造成的损伤是不一样的。即在语音的时间-频率表示中(语谱图)，有的区域受噪声污染的程度小一点，从这些区域提取出来的特征有比较高的置信度，在识别过程中具有比较高的鉴别能力；与之相反，那些从受

其中 $Y(k, t)$ 和 $e^{j\varphi(k, t)}$ 分别表示频谱域第 k 个频点的幅度和相位, N_{fft} 是 FFT 变换的点数。

求功率谱由于语音的短时谱幅度对感知语音起主导作用, 而短时相位相对来说在听觉上并不很重要, 因此可以计算功率谱幅度, 而忽略相位的影响, 表达形式为:

$$Y_p(k, t) = |Y(k, t)|^2, k = 1, \dots, N_{fft} \quad (15)$$

Mel-Scaled 滤波器组 Mel 频段划分是在对听觉模型的研究基础上提出的。Mel-Scaled 频率 f_{mel} 与线性频率 f_{Hz} 的关系为:

$$f_{mel} = 1127 \ln(1 + \frac{f_{Hz}}{700}) \quad (16)$$

Mel 滤波器组如图 5 所示。首先利用式(16)将线性频率, 即 FFT 变换后的频率变换到 Mel 频率上, 并在 Mel 频率上进行均匀分段。M 表示功率谱域上 Mel-Scaled 滤波器组的个数, 也即 Mel 频率上的分段个数:

$$Mel_m = m \times 1127 \ln(1 + \frac{F_s/2}{700}) / M, m = 1, \dots, M \quad (17)$$

其中 Mel_m 表示第 m 个 Mel 分段频率, F_s 是信号的采样频率, 然后将 Mel 分段频率映射回线性频率:

$$Lin_m = (\exp(Mel_m / 1127) - 1) \times 700, m = 1, \dots, M \quad (18)$$

其中 Lin_m 表示第 m 个 Mel 分段频率对应的线性频率, 计算 Mel 滤波器组在各线性频点上的抽头系数:

$$H_m(k) = \begin{cases} \frac{f_k - Lin_{m-1}}{Lin_m - Lin_{m-1}} & Lin_{m-1} \leq f_k \leq Lin_m \\ 0 & \text{else} \\ \frac{f_k - Lin_m}{Lin_{m+1} - Lin_m} & Lin_m \leq f_k \leq Lin_{m+1} \end{cases}, k = 1, \dots, N_{fft}/2, m = 1, \dots, M \quad (19)$$

其中 $H_m(k)$ 表示第 m 个 Mel 滤波器在第 k 个线性频点上的抽头系数, f_k 表示第 k 个频点的频率值:

$$f_k = k \times F_s / N_{fft}, k = 1, \dots, N_{fft} \quad (20)$$

提取的 Mel 谱特征为:

$$MBank(m, t) = \sum_{k=1}^{N_{fft}/2} H_m(k) \times Y_p(k, t), m = 1, \dots, M \quad (21)$$

其中 $MBank(m, t)$ 表示提取的第 t 帧语音的第 m 维 Mel 谱特征。

对数谱表示考虑到人的听觉特性, 如对声音响度的感觉是与声强的对数值呈线性关系的,

噪声污染大的区域提取出来的特征，将对识别造成干扰，是识别率下降的主要原因。

在基于 HMM 的统计识别方法中，获得和利用的先验知识越多，识别的结果就越准确。特征加权算法利用了噪声在不同时间-频率区域对语音的损伤程度信息，有效的提高了噪声环境下的识别性能，即根据短时段语音各维特征提取空间的局部信噪比，给出特征的置信度估计，即权重，并对识别算法进行修改，将权重信息代入识别过程。

特征加权算法需要解决以下两个问题：

- 1.如何估计特征的置信度，并给出权重计算公式。
- 2.如何将特征加权过程嵌入到基于 HMM 的识别框架中。

从 MFCC 特征的提取过程(图 4)中，可以看出在进行 DCT 变换之前，我们称之为对数谱特征，它的每一维数据都和含噪语音在当前短时段的某个局部频率区间相联系。因此各维对数谱特征的置信度可以通过此区间的局部信噪比进行估计。

在语音增强技术中，基于语音短时谱幅度 (STSA: Short Time Spectral Amplitude) 估计的方法利用了语音在听觉方面的一个重要特性，即语音短时频谱对感知语音起决定性的作用，其中语音的短时谱幅度又是起主导作用的，而语音的短时相位相对来说在听觉上并不很重要。因此，基于 STSA 估计的语音增强方法一般只增强语音的 STSA，而直接把带噪语音的相位作为增强语音的相位。图 6 给出了该类方法的一般框图。

噪声语音与纯净语音在时域的叠加为含噪语音，即含噪语音可以表示为

$$y(n) = x(n) + d(n) \quad (28)$$

其中 $x(n)$ 是纯净语音， $d(n)$ 是加性背景噪声，且两者互不相关。识别和增强处理都需要将语音按短时段进行划分，经过分帧和加窗处理后，公式(4.1)表示为：

$$y_w(n, t) = x_w(n, t) + d_w(n, t), 1 \leq n \leq N \quad (29)$$

其中 N 为语音帧的长度， t 是帧序号。 $d(n, t)$ ， $x(n, t)$ ， $y(n, t)$ 的短时离散频谱幅度和短时离散功率谱幅度分别用 $D(k, t)$ ， $X(k, t)$ ， $Y(k, t)$ 和 $D_p(k, t)$ ， $X_p(k, t)$ ， $Y_p(k, t)$ 表示，其中 $1 \leq k \leq N_{\text{fft}}$ 表示各个频点， N_{fft} 为一帧快速傅立叶变换(FFT)的长度。

基于 STSA 估计的语音增强方法有一个通用的增强估计公式。

$$\hat{X}(k, t) = G(k, t)Y(k, t), 1 \leq k \leq N_{\text{fft}}/2 + 1 \quad (30)$$

其中 $G(k, t)$ 称为第 t 帧中第 k 个频点的增益系数，在不同的增强方法中有不同的函数表达形式。 $\hat{X}(k, t)$ 表示 $X(k, t)$ 的估计，即增强后语音的短时谱幅度。MMSE 方法的核心是计算纯净语音短时谱幅度 $X(k, t)$ 的最小均方误差估计，在语音和噪声频谱的高斯分布假设下，增益系数可以表示为：

$$G(k, t) = \frac{1}{2} \sqrt{\frac{\pi \zeta(k, t)}{\gamma(k, t)(1 + \zeta(k, t))}} \Psi(-0.5; 1; -\frac{\gamma(k, t)\zeta(k, t)}{1 + \zeta(k, t)}), 1 \leq k \leq N_{\text{fft}}/2 + 1 \quad (31)$$

其中 $\zeta(k, t)$ ， $\gamma(k, t)$ 分别称为先验信噪比和后验信噪比， $\Psi(a_1, a_2, a_3)$ 为合流超几何函数，可

$$\log(p(y_i | q(t) = i)) = \log N(y_i; \mu_i, \Sigma_i) = -\sum_{r=1}^R \log(\sqrt{2\pi}\sigma_{ir}) - \sum_{r=1}^R \frac{(y_{ir} - \mu_{ir})^2}{\sigma_{ir}^2} \quad (35)$$

为了便于表述特征加权算法, 令 μ^c 表示高斯分布的均值矢量, Σ^c 表示方差矩阵, 上标 c 表示倒谱域。由于倒谱特征各维之间的近似不相关特性, 可以令 Σ^c 为对角矩阵。 R 维特征矢量 y_i^c 在此高斯模型下的对数概率似然值为:

$$\log N(y_i^c, u^c, \Sigma^c) = c(\Sigma^c) - \frac{1}{2} d^{cT} \Sigma^{c-1} d^c, d^c = y_i^c - u^c \quad (36)$$

其中 $c(\Sigma^c)$ 表示与 y_i^c 无关的常数项, 可以看出 $c(\Sigma^c)$ 对应于式(4-8)中的 $-\sum_{r=1}^R \log(\sqrt{2\pi}\sigma_{ir})$, 而

$$\frac{1}{2} d^{cT} \Sigma^{c-1} d^c \text{ 对应于 } \sum_{r=1}^R \frac{(y_{ir} - \mu_{ir})^2}{\sigma_{ir}^2}。$$

在对数谱域进行特征加权非常直观, 它的公式如下:

$$\log^* N(y_i^l, u^l, \Sigma^l) = c(\Sigma^l) - \frac{1}{2} d^{lT} W^T (\Sigma^l)^{-1} W d^l, d^l = y_i^l - u^l \quad (37)$$

其中 y_i^l 表示第 i 帧 M 维对数谱特征, 权重矩阵 $W = \text{diag}\{w_1(t), w_2(t), \dots, w_m(t)\dots\}$, 元素 $w_m(t)$ 是第 m 维对数谱特征的权重, 上标 l 表示对数谱域。

综合公式(36)和(37), 可以得到倒谱域上的特征加权对数似然计算公式:

$$\log^* N(y_i^c, u^c, \Sigma^c) = c(\Sigma^c) - \frac{1}{2} d^{lT} W^T Tr^T (\Sigma^c)^{-1} Tr W d^l, d^l = Tr^{-1}(y_i^c - u^c) \quad (38)$$

公式(38)的意义可以表述为: 首先在倒谱域计算倒谱特征和状态均值的差值矢量, 将其变换到对数谱域进行加权, 然后再变换回倒谱特征进行识别。其中矩阵 Tr 表示图 4 中的 DCT 变换和倒谱加权, 即从对数谱特征到 MFCC 特征的线性变换。

$$Tr = DCT \times \text{Diag}\{1, 1 + \frac{L}{2} \sin(\frac{\pi}{L}), \dots, 1 + \frac{L}{2} \sin(\frac{\pi(R-1)}{L}), 0, \dots, 0\} \quad (39)$$

即 DCT 矩阵和倒谱加权对角阵的乘积, 倒谱加权对角阵的对角元素的前 R 维是倒谱加权系数, 而后面的 $M-R$ 维为 0。 Tr^{-1} 是它的逆变换, 可以表示为:

$$Tr^{-1} = \text{Diag}\{1, \frac{1}{1 + \frac{L}{2} \sin(\frac{\pi}{L})}, \dots, \frac{1}{1 + \frac{L}{2} \sin(\frac{\pi(R-1)}{L})}, 0, \dots, 0\} \times DCT^{-1} \quad (40)$$

具体来说, 由于我们识别时采用的 MFCC 特征的维数 R 小于对数谱特征维数 M , 在进行这种特征转换时, 我们对 MFCC 特征增加维数, 增加的各维特征数据用 0 代替; 在识别过程中, 仍采用 R 维的 MFCC 特征。

在基于 HMM 的识别框架下, 噪声所带来的测试和识别环境的失配可以映射到三个空间, 即信号、特征和模型空间。抗噪声语音识别的方法也是从这三个方面考虑的。本发明对这三个空间的抗噪声语音识别技术进行融合, 提出多空间信号处理的抗噪声语音识别方案, 期望在低信噪比加性噪声环境下进一步提高识别精度。

以利用级数求和计算:

$$\Psi(a_1, a_2, a_3) = 1 + \frac{a_1}{a_2} \frac{a_3}{1} + \frac{a_1(a_1+1)}{a_2(a_2+1)} \frac{a_3^2}{2!} + \dots \quad (32)$$

其中, $a_1 = -0.5$, $a_2 = 1$, $a_3 = -\frac{\gamma(k,t)\zeta(k,t)}{1+\zeta(k,t)}$ 。

可以看出, 代表局部信噪比的 $\zeta(k,t)$ 和 $\gamma(k,t)$ 越大, 增益系数 $G(k,t)$ 也越大, 反之亦然。因此 $G(k,t)$ 可以作为局部信噪比的度量, 用于特征权重的计算。

由于特征权重是在对数谱域和每一维特征相关的, 因此我们从对数谱特征的提取过程(图4)得到借鉴, 计算特征权重。

$$w_m(t) = \sum_{k=1}^{N_p/2} G(k,t) H_m(k) / \sum_{k=1}^{N_p/2} H_m(k), 1 \leq m \leq M \quad (33)$$

然后进行规范化, 使

$$\sum_{m=1}^M w_m(t) = 1 \quad (34)$$

这里 $H_m(k)$ 是图5中功率谱域中第 m 个三角滤波器在第 k 个频谱分量上的系数, 见公式(19), 而 $w_m(t)$ 表示第 t 帧语音提取的第 m 维对数谱特征的权重。 M 是 Mel 滤波器的个数, 也即对数谱特征的维数。

图7给出了在 0dB 加性高斯白噪声环境下, 某帧浊音段语音 26 维对数谱特征的失配情况(图7a)和采用上述方法得到的特征权重(图7b)。可以看出失配越大, 权重越小, 反之亦然。特别是特征权重在反映语音内容信息的两个共振峰频率附近有明显的峰值, 突出这部分信息将有利于提高语音的识别精度。

在语音无声段, 采用上述方法得到的特征权重与特征的实际失配情况不符, 我们不对其进行特征加权, 即令各维特征的权重为 1。

虽然我们是在对数谱域对特征进行加权, 但由于对数谱特征进行识别的性能不如倒谱特征, 而且倒谱特征具有维数低, 各维数据近似不相关等特性, 可以简化语音模型和减少识别运算量, 因此我们的特征加权识别算法中, 依然采用倒谱域的 MFCC 特征。

识别程序采用 Viterbi 译码算法, 即寻找最大对数似然输出状态序列:

$$\begin{aligned} \delta_t(i) &= \max_{q(1)q(2)\dots q(t-1)} \log[p(q(1)q(2)\dots q(t-1), q(t)=i, y_1 y_2 \dots y_t | \lambda)] \\ &= \max_{1 \leq j \leq N} [\delta_{t-1}(i) + \log(a_{ij})] + \log(p(y_t | q(t)=j)) \end{aligned} \quad (4/6)$$

因此特征加权算法的核心在于获得针对特征失配, 具有鲁棒性的对数似然计算公式。将公式(3)代入对数似然计算公式

含噪语音经过 MMSE 处理后，得到的纯净语音的估计值与真值之间存在一定的误差，我们称之为剩余噪声，可以表示为：

$$\hat{d}(n) = \hat{x}(n) - x(n) \quad (41)$$

其中 $\hat{d}(n)$ 和 $x(n)$ 分别表示剩余噪声和纯净语音在第 n 个样点上的值， $\hat{x}(n)$ 表示 $x(n)$ 的估计值，为了消除这部分剩余噪声带来的训练和测试环境的失配，我们考虑在模型空间进行纯净语音训练模型的噪声补偿。MMSE 增强后的剩余噪声保持了一定的准平稳特性，我们可以用一个单高斯状态分布的 HMM 来描述，模型空间我们采用了 Log-Add 方法，只对纯净语音训练模型的状态均值进行补偿，在不影响识别率的前提下，极大地降低了计算复杂度，同时不需要对剩余噪声进行模型训练，而只需要估计剩余噪声的特征均值，这些都有利于方案的实时实现。

由于剩余噪声存在于各个语音帧，而语音仅存在于非噪声帧，所以对于噪声帧来说， $\hat{D}(k, t) = \hat{X}(k, t)$ ，其中 $\hat{X}(k, t)$ 表示纯净语音在第 t 帧中第 k 个频点上的谱幅度的估计值，而 $\hat{D}(k, t)$ 是剩余噪声在第 t 帧中第 k 个频点上的谱幅度，即剩余噪声的短时谱幅度在各噪声帧中等于增强后语音的短时谱幅度。利用信号空间 MMSE 语音增强时获得的噪声帧检测信息，对所有从这些增强后的噪声帧中提取的 MFCC 特征求均值，便可以获得用于 Log-Add 模型补偿的剩余噪声的特征均值。

$$\mu_{nr}^c = \frac{\sum_{\text{第 } t \text{ 帧为噪声帧}} c(r, t)}{\text{噪声帧的个数}} \quad (42)$$

本发明提出的多空间融合抗噪声语音识别技术可以简述如下：

选择 MMSE 法对含噪语音进行前端语音增强。首先是因为它的计算复杂度低，可以实时地处理；其次是因为它在处理过程中提供的辅助信息(增益系数)可以比较准确的估计对数谱特征的权重，反映各维特征的失配情况；最后，MMSE 增强处理对语音的损伤比较小，处理后的剩余噪声保持原有的准平稳特性，有利于后面的模型补偿。

选择前面提出的特征加权算法，利用信号空间增强算法获得的谱幅度增益值估计对数谱特征的权重，并将此权重信息引入识别过程。

选择算法复杂度较低的对数谱相加(Log-Add)补偿算法，即将纯净语音模型和噪声模型的 MFCC 特征均值分量在对数谱域相加，从而得到补偿后的含噪语音模型的对数谱均值。与经典的并行模型补偿(PMC)算法相比，它只对模型的均值，而不对方差进行补偿，计算量远远小于 PMC，但可以达到基本相同的识别精度。此外 Log-Add 算法不仅可以补偿静态 MFCC 的均值，而且可以补偿动态与高阶 MFCC。

$$\hat{\mu}_m^l = \mu_m^l + \log(1 + \exp(\mu_{nm}^l - \mu_m^l)) \quad (43)$$

$\Delta\mu^c$ ：纯净语音训练得到的模型状态在 MFCC 倒谱域下的动态特征均值；

(3).分帧、加窗：

若采样后的原始语音为 $y(n)$ ，汉明(hamming)窗在第 n 个采样点上的系数：

$$h(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), n = 1, \dots, N$$

N 等于帧长，则分帧后的原始语音信号为：

$$y(n, t) = y\left(\frac{N \times (t-1)}{2} + n\right), n = 1, \dots, N$$

t 表示帧号，加上汉明窗后的原始语音信号为：

$$y_w(n, t) = y(n, t) \times h(n), n = 1, \dots, N$$

(4). 快速傅立叶变换 FFT：

由于语音短时频谱对感知语音起决定性的作用，利用 FFT 逐帧将语音变换到频谱域：

$$\bar{Y}(k, t) = Y(k, t) e^{j\angle \bar{Y}(k, t)} = FFT\{y_w(n, t)\}, k = 1, \dots, N_{fft}$$

N_{fft} 是 FFT 变换的点数。

(5).噪声帧检测和噪声谱幅度估计：

(5.1).设定前 10 帧起始段含噪语音为噪声帧，输入当前第 t 帧含噪语音的短时谱幅度：

(5.2).若当前帧为起始段噪声帧,则前 t 帧噪声功率谱幅度的估计值为：

$$\tilde{D}_p(k, t) = \left[\sum_{s=1}^t Y(k, s) / t \right]^2$$

并在当前帧为第 10 帧时输出起始段噪声谱幅度的估计值：

$$N(k) = \sum_{s=1}^{10} Y(k, s) / 10$$

计算用于区分噪声帧和含噪语音帧的判决门限 χ ：

$$\chi = \text{Max}_{t=1,2,\dots,10} \left\{ \sum_{k=1}^{N_{fft}/2+1} \text{Pow}[Y(k, t) / N(k), 5] \right\}$$

(5.3).若当前帧不是起始段噪声帧，则当前帧 t 的判决值：

$$\rho = \sum_{k=1}^{N_{fft}/2+1} \text{Pow}[Y(k, t) / N(k), 5]$$

(5.3.1)若 $\rho < \chi$ ，则判决为含噪语音中的噪声帧，其噪声功率谱幅度估计值为：

$$\tilde{D}_p(k, t) = 0.98 \times \tilde{D}_p(k, t-1) + 0.02 \times Y_p(k, t)$$

并输出：

(5.3.2).若 $\rho \geq \chi$ ，则判决为非噪声帧，即含有噪声的语音帧，其噪声功率谱幅度为：

$$\tilde{D}_p(k, t) = \tilde{D}_p(k, t-1)$$

$$\Delta \hat{\mu}_m^l = \frac{\Delta \mu_m^l}{1 + \exp(\mu_{nm}^l - \mu_m^l)} \quad (44)$$

其中, $\hat{\mu}^l$ 和 $\Delta \hat{\mu}^l$ 分别表示补偿后的模型在对数谱域的静态和动态状态均值; μ^l 和 $\Delta \mu^l$ 表示纯净语音训练得到的模型在对数谱域的静态和动态状态均值; μ_n^l 是剩余噪声的特征均值; 上标 l 表示对数谱域; 下标 m 代表第 m 维特征。

由于我们是在 MFCC 倒谱域得到纯净语音训练模型的状态均值和剩余噪声特征均值, 而实际的模型补偿是在对数谱域进行, 这同样需要进行对数谱特征和 MFCC 特征之间的转换, 与特征加权算法相同, 即对低维的 MFCC 特征增加维数, 并利用线性变换 Tr 和 Tr^{-1} 进行转换。即: $\mu_n^l = Tr^{-1} \mu_n^c$, $\mu^l = Tr^{-1} \mu^c$, $\Delta \mu^l = Tr^{-1} \Delta \mu^c$, $\hat{\mu}^c = Tr \hat{\mu}^l$, $\Delta \hat{\mu}^c = Tr \Delta \hat{\mu}^l$ 。其中, $\hat{\mu}^c$ 、 $\Delta \hat{\mu}^c$ 、 μ^c 、 $\Delta \mu^c$ 和 μ_n^c 分别表示 $\hat{\mu}^l$ 、 $\Delta \hat{\mu}^l$ 、 μ^l 、 $\Delta \mu^l$ 和 μ_n^l 对应的 MFCC 倒谱特征, 上标 c 表示 MFCC 倒谱域。最后得到剩余噪声补偿后的语音模型在各状态下的静态和动态 MFCC 均值。

为了简便计算, 我们用单高斯状态分布的 HMM 描述剩余噪声模型, 模型补偿时只需要剩余噪声的特征均值, 因此不需要对噪声模型进行离线训练, 有利于识别方案的实时处理。

MMSE-FW-LA 方案的算法流程如图 8 所示:

1. 首先输入含噪语音和纯净语音训练得到的语音模型, 对含噪语音进行分帧和加窗, 并做 FFT 变换到频域。
2. 进行语音间歇, 即噪声段检测, 并估计噪声的功率谱幅度。
3. 用 MMSE 法估计纯净语音的短时谱幅度, 并保留谱幅度增益系数。
4. 利用上一步得到的谱幅度增益系数计算对数谱特征的权重。
5. 利用第三步得到的增强语音的短时谱幅度, 即纯净语音短时谱幅度的估计值提取 MFCC 特征。
6. 利用第二步得到的无声段划分信息, 和上一步得到的增强语音的 MFCC 特征, 计算剩余噪声的 MFCC 特征均值。
7. 在模型空间, 用 Log-Add 法对纯净语音训练得到的语音模型做剩余噪声补偿。这里利用了上一步得到的剩余噪声的 MFCC 特征均值。
8. 将第五步得到的增强语音的 MFCC 特征参数、上一步得到的剩余噪声补偿后的语音模型以及第四步得到的对数谱特征权重输入基于特征加权的识别解码器。
9. 得到识别结果。

本发明的特征在于: 它依次含有以下步骤:

(1). 初始化 Mel 滤波器组在各线性频点 k 上的抽头系数 $H_m(k)$, 以及对数谱特征与 MFCC(Mel 频段倒谱系数)特征的转换矩阵 Tr 和 Tr^{-1} : 其中 $k=1, 2, \dots, N_p/2$, N_p 是 FFT 的频点数; $m=1, 2, \dots, M$, M 是 Mel 滤波器的个数。

(2). 输入含噪语音和纯净语音经训练得到的模型参数:

μ^c : 纯净语音训练得到的模型状态在 MFCC 倒谱域下的静态特征均值;

并输出:

(6).用取决于先验信噪比 ζ 和后验信噪比 γ 的谱幅度增益系数 $G(k,t)$ 计算纯净语音短时谱幅度的估计值, 以及相应的第 t 帧第 m 个对数谱特征的权重 $w_m(t)$:

(6.1).输入当前第 t 帧含噪语音的短时谱幅度;

(6.2).计算当前帧 t 第 k 个频点的后验信噪比 $\gamma(k,t)=Y_p(k,t)/\bar{D}_p(k,t)$, $Y_p(k,t)$ 为含噪语音的功率谱幅度, $\bar{D}_p(k,t)$ 为估计的噪声功率谱幅度。

(6.2.1).如果当前帧 $t=1$, 则初始化当前帧第 k 个频点的先验信噪比为 $\zeta(k,t)=0.1$;

(6.2.2).如果当前帧 $t>1$, 则利用上一帧的先验和当前帧的后验信噪比, 通过滑动平均估计得到当前帧第 k 个频点的先验信噪比:

$$\zeta(k,t)=0.98 \times \zeta(k,t-1)+0.02 \times [\gamma(k,t)-1]$$

(6.3).当前帧 t 第 k 个频点的谱幅度增益系数:

$$G(k,t)=\frac{1}{2} \sqrt{\frac{\pi \zeta(k,t)}{\gamma(k,t)(1+\zeta(k,t))}} \Psi(-0.5;1;-\frac{\gamma(k,t)\zeta(k,t)}{1+\zeta(k,t)})$$

利用级数求和, 计算得到:

$$\Psi(a_1,a_2,a_3)=1+\frac{a_1}{a_2} \frac{a_3}{1} + \frac{a_1(a_1+1)}{a_2(a_2+1)} \frac{a_3^2}{2!} + \dots$$

$$\text{其中 } a_1=-0.5, \quad a_2=1, \quad a_3=-\frac{\gamma(k,t)\zeta(k,t)}{1+\zeta(k,t)}$$

(6.4).相应的纯净语音短时谱幅度的估计值为:

$$\hat{X}(k,t)=G(k,t)Y(k,t)$$

(6.5).重新计算当前帧第 k 个频点的先验信噪比:

$$\zeta(k,t)=|\hat{X}(k,t)|^2/\bar{D}_p(k,t)$$

(6.6).计算完当前帧 t 第 k 个频点($1 \leq k \leq N_{\text{fft}}/2+1$)的 $G(k,t)$ 、 $\hat{X}(k,t)$ 和 $\zeta(k,t)$ 值。

(6.7).计算当前帧 t 第 m 个对数谱特征的权重:

$$w_m(t)=\sum_{k=1}^{N_{\text{fft}}/2} G(k,t)H_m(k)/\sum_{k=1}^{N_{\text{fft}}/2} H_m(k)$$

(6.8).计算当前帧共 M 个对数谱特征的权重, M 是对数谱特征的维数。

(6.9).计算完 $t=1,2,\dots,T$ 各帧中的 $\hat{X}(k,t)$ 和 $w_m(t)$;

(6.10).输出所有相应的纯净语音短时谱幅度估计值 $\hat{X}(k,t)$ 和对数谱特征的权重 $w_m(t)$;

(7).MFCC 特征提取

(7.1).输入纯净语音短时谱幅度估计值 $\hat{X}(k,t)$;

(7.2).计算功率谱: $\hat{X}_p(k,t)=|\hat{X}(k,t)|^2, k=1,\dots,N_{\text{fft}}$;

(7.3).Mel 滤波:

$$MBank(m,t) = \sum_{k=1}^{N_p/2} H_m(k) \times \hat{X}_p(k,t), m=1,\dots,M$$

(7.4).对数谱特征: $FBank(m,t) = \log(MBank(m,t)), m=1,\dots,M$

(7.5).DCT 倒谱表示:

$$\tilde{c}(r,t) = \alpha(r) \sum_{m=1}^M FBank(m,t) \cos\left(\frac{\pi(2m-1)(r-1)}{2M}\right), r=1,\dots,M$$

其中 $\alpha(1) = \sqrt{\frac{1}{M}}, \alpha(r) = \sqrt{\frac{2}{M}}, r=2,\dots,M$, 并取前 R 维特征矢量

(7.6).倒谱加权:

$$c(r,t) = lifter(r) \times \tilde{c}(r,t), r=1,\dots,R$$

其中 $lifter(r) = 1 + \frac{L}{2} \sin\left(\frac{\pi(r-1)}{L}\right), r=1,\dots,R$, L 为加权滤波器宽度;

(7.7).计算动态系数:

$$\Delta c(r,t) = \sum_{\Delta t=-2}^2 \Delta t c(r,t+\Delta t) / 10, \Delta t \text{ 表示帧间距};$$

(7.8).输出 $c(r,t)$ 和 $\Delta c(r,t)$;

(8).判断待识别语句是否输入完毕 $t=T$?

(9).若判断为待识别语句已经输入完毕,则计算噪声帧,即剩余噪声的静态 MFCC 特征平均值,剩余噪声的定义如下:

$$\hat{d}(n) = \hat{x}(n) - x(n)$$

其中 $x(n)$ 表示纯净语音在第 n 个样点上的值, $\hat{x}(n)$ 表示 $x(n)$ 增强后的估计值。由于剩余噪声存在于各个语音帧,而语音仅存在于非噪声帧,所以对于噪声帧来说, $\hat{D}(k,t) = \hat{X}(k,t)$, 即剩余噪声的短时谱幅度在各噪声帧中等于增强后语音的短时谱幅度,我们可以利用下式计算剩余噪声的静态 MFCC 特征均值:

$$\mu_{nr}^c = \frac{\sum_{\text{第 } t \text{ 帧为噪声帧}} c(r,t)}{\text{噪声帧的个数}}$$

其中噪声帧包括起始段 10 帧和后面判决的噪声帧, $r=1,2,\dots,R$ 。

(10).Log-Add 对数谱相加模型补偿:

(10.1).输入剩余噪声的 MFCC 特征均值并转换到对数谱域 $\mu_n^l = Tr^{-1} \mu_n^c$;

(10.2).输入纯净语音训练模型的状态均值,并转换到对数谱域 $\mu^l = Tr^{-1} \mu^c$,
 $\Delta \mu^l = Tr^{-1} \Delta \mu^c$;

(10.3).Log-Add 模型补偿:

$$\hat{\mu}_m^l = \mu_m^l + \log(1 + \exp(\mu_{nm}^l - \mu_m^l)), m = 1, 2, \dots, M$$

$$\Delta \hat{\mu}_m^l = \frac{\Delta \mu_m^l}{1 + \exp(\mu_{nm}^l - \mu_m^l)}$$

(10.4).把补偿的模型状态转换到 MFCC 倒谱域 $\hat{\mu}^c = Tr \hat{\mu}^l$, $\Delta \hat{\mu}^c = Tr \Delta \hat{\mu}^l$;

(10.5).当状态输入完毕, 输出剩余噪声补偿后的语音模型;

(11).特征加权的维特比识别译码:

(11.1).输入剩余噪声补偿后的语音模型、增强语音当前帧 MFCC 特征 y_i^c 、对数谱特征权重 $w_m(t)$;

(11.2).计算观测帧在候选状态下的对数概率似然值:

(11.2.1).在 MFCC 倒谱域计算 MFCC 特征与可选状态的状态均值的矢量差:

$$d^c = y_i^c - u^c;$$

(11.2.2).把差矢量变换到对数谱特征域: $d^l = Tr^{-1} d^c$;

(11.2.3).在对数谱域进行加权, 并变换回 MFCC 倒谱域 $\bar{d}^c = Tr W d^l$;

(11.2.4).计算对数概率似然值:

$$\log(p(y_i^c | q(t) = i)) = C(\Sigma^c) - \frac{1}{2} \bar{d}^{cT} (\Sigma^c)^{-1} \bar{d}^c$$

其中 Σ^c 为倒谱域的状态方差矩阵, 且为对角阵 $\Sigma^c = \text{Diag}\{\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{iR}\}$, c 表示倒谱

域, i 表示状态; $C(\Sigma^c)$ 表示与 y_i^c 无关的常数项, 对应 $-\sum_{r=1}^R \log(\sqrt{2\pi}\sigma_{ir})$, R 是倒谱

特征的维数。

(11.3).初始化 Viterbi 译码后, 再迭代, 计算完 $t=1, 2, \dots, T$ 帧;

(11.4).计算最大概率 $p^* = \max_{1 \leq i \leq N} [\delta_T(i)]$ 和最佳路径的终止状态: $\hat{q}(T) = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$;

(11.5)通过回溯依次输出最佳路径上的其他状态: $\hat{q}(t) = \varphi_{t+1}(\hat{q}(t+1)), t = T-1, \dots, 1$;

(12). 输出识别结果, 结束。

使用证明: 它达到了预期目标。

附图说明使用证明: 它达到了预期目标。

附图说明

图 1: HMM 在语音识别中的应用。

图 2: 环境噪声模型。

图 3: 训练和识别的失配。

图 4: MFCC 特征提取过程。

图 5: Mel 滤波器组构造图。

图 6: 基于 STSA 估计的语音增强框图。

图 7: 信噪比 0dB 白噪声环境下对数谱特征失配和权重示意图:

a: 26 维对数谱矢量;

b: 26 维对数谱矢量权重。

图 8: MMSE-FW-LA 方案算法流程图。

图 9: MMSE-LA 方案主程序流程图。

图 10: MMSE-FW-LA 方案主程序流程图。

图 11: 噪声段检测/噪声功率谱幅度估计核心程序流程图

图 12: 语音增强和特征权重计算核心程序流程图。

图 13: MFCC 特征提取算法框图。

图 14: Log-Add 模型补偿核心流程图。

图 15: 特征加权的维特比识别译码核心程序流程图。

图 16: 低信噪比白噪声环境下, 前端 MMSE 增强、特征加权和 Log-Add 模型补偿的抗噪声识别性能比较。

图 17: 低信噪比白噪声环境下, 特征加权分别和前端 MMSE 增强、Log-Add 模型补偿融合后的抗噪声识别精度比较。

图 18: 低信噪比白噪声环境下, MMSE-FW-LA 与 MMSE-LA 方案的抗噪声识别性能比较。

图 19: 低信噪比汽车噪声环境下, 前端 MMSE 增强、特征加权和 Log-Add 模型补偿的抗噪声识别性能比较。

图 20: 低信噪比汽车噪声环境下, 特征加权分别和前端 MMSE 增强、Log-Add 模型补偿融合后的抗噪声识别精度比较。

图 21: 低信噪比汽车噪声环境下, MMSE-FW-LA 与 MMSE-LA 方案的抗噪声识别性能比较。

从图 9、10 可以看出, MMSE-FW-LA 和 MMSE-LA 方案的主程序流程基本相同, 只是多了一个对数谱特征权重计算模块, 并且识别时采用特征加权的维特比译码器。整个算法流程包括五个核心模块: 噪声帧检测和噪声功率谱幅度估计模块、MMSE 语音增强和对数谱特征权重估计、MFCC 特征提取、Log-Add 模型补偿和特征加权的维特比译码算法。

图 11 给出了噪声帧判决和噪声的短时谱幅度估计模块的流程图, 输入为含噪语音当前帧的短时谱幅度, 输出是噪声帧的判决结果和经过当前帧估计更新后的噪声功率谱幅度。噪声帧检测采用了基于能量的检测方法。

由于待识别的含噪语音开头总有一个无声段, 因此我们将前 10 帧语音判决为噪声帧, 噪声的功率谱幅度的估计值为:

$$\tilde{D}_p(k, t) = \left[\sum_{s=1}^t Y(k, s) / t \right]^2 \quad (45)$$

其中 $Y(k, s)$ 表示输入的含噪语音短时谱幅度，然后计算判决门限：

$$\chi = \text{Max}_{t=1,2,\dots,10} \left\{ \sum_{k=1}^{N_g/2+1} \text{Pow}[Y(k, t) / N(k), 5] \right\} \quad (46)$$

其中 $N(k) = \sum_{s=1}^{10} Y(k, s) / 10$ 表示粗估的噪声谱幅度，函数 $\text{Pow}(x_1, x_2) = x_1^{x_2}$ 。

从第 11 帧开始，需要进行噪声帧检测判决，首先计算判决参数：

$$\rho = \sum_{k=1}^{N_g/2+1} \text{Pow}[Y(k, t) / N(k), 5] \quad (47)$$

如果 $\rho < \chi$ ，判决为噪声帧，此时需要重新估计噪声功率谱幅度：

$$\tilde{D}_p(k, t) = 0.98 \times \tilde{D}_p(k, t-1) + 0.02 \times Y_p(k, t) \quad (48)$$

即进行系数为 0.98 的平滑估计。如果 $\rho \geq \chi$ ，则不做噪声功率谱幅度的重新估计：

$$\tilde{D}_p(k, t) = \tilde{D}_p(k, t-1) \quad (49)$$

图 12 给出 MMSE 语音增强和对数谱特征权重估计模块的流程图，其输入是含噪语音当前帧的短时谱幅度，输出为增强后语音的短时谱幅度，即纯净语音的短时谱幅度的估计，和对数谱特征权重。由于计算短时谱幅度增益系数需要计算含噪语音在当前频点的先验和后验信噪比，如式(31)所示：

$$G(k, t) = \frac{1}{2} \sqrt{\frac{\pi \zeta(k, t)}{\gamma(k, t)(1 + \zeta(k, t))}} \Psi(-0.5; 1; -\frac{\gamma(k, t) \zeta(k, t)}{1 + \zeta(k, t)}) \quad (31)$$

在实际运算中，先验信噪比可以通过滑动平均估计得到：

$$\zeta(k, t) = 0.98 \times \zeta(k, t-1) + 0.02 \times [\gamma(k, t) - 1] \quad (50)$$

后验信噪比可以直接计算得到：

$$\gamma(k, t) = Y_p(k, t) / \tilde{D}_p(k, t) \quad (51)$$

$\tilde{D}_p(k, t)$ 为估计的噪声功率谱幅度，参见图 11。

图 13 给出了 MFCC 特征提取模块的程序流程图，输入为增强语音的谱幅度值，输出为增强语音的 MFCC 特征参数。

图 14 给出了 Log-Add 模型补偿核心流程图，输入为纯净语音训练得到的语音模型和剩余噪声的 MFCC 特征均值，输出为剩余噪声补偿后的语音模型。

图 15 给出了特征加权的维特比识别译码核心程序流程图，输入为经过 MMSE 增强的语音特征、对数谱特征权重和剩余噪声补偿后的语音模型，输出为识别结果。

本发明内容主要讨论强背景噪声环境下的抗噪声语音识别，识别系统针对非特定人连续语音数字串，具体的实验描述如下：

基线系统(BaseLine)

为了便于进行实验结果的比较，我们首先搭建了一个连续语音识别系统，它由三个模块组成：MFCC 特征提取、训练模块和识别模块。

基线系统采用的特征是 26 维的 MFCC_0_D 特征。其中 MFCC 表示除 $c(l,t)$ 之外的静态倒谱，0 表示反映语音能量信息的 $c(l,t)$ 谱，D 表示根据静态倒谱或 MFCC_0 求出的一阶倒谱。MFCC 特征参数设置如下：

语音短时帧的长度为 20ms，即 $N=320$ ；帧交叠为 10ms，即 160 个采样点。短时帧 FFT 的点数 $N_f=512$ 。

Mel 滤波器的个数为 $M=26$ 。

静态的 MFCC 参数个数为 $R=13$ 。

倒谱加权系数 $L=22$ 。

由于是小词汇量连续语音识别，基线系统采用 12 个连续，状态无跨越由左到右的 HMM 字模型('one'~'nine', 'oh', 'zero' and 'sil')，每个模型有 8 个状态，各个状态的特征概率分布用单个对角化多维高斯分布来近似。

语音数据库

实验的训练和测试语音数据库为 TI-Digits。TI-Digits 由 Texas Instruments 公司设计，用来训练和测试非特定人英文数字串语音识别系统，共有 326 人(111 个成年男性，114 个成年女性，50 个男孩，51 个女孩)，每人 77 个数字串发音。实验训练使用 TIDigit 库中 15 个说话人的 500 句话，识别测试使用库中与训练无关的 4 个人的 100 句话。语音数据的采样率为 16KHz，采样比特为 16bit。

噪声数据库

实验用的噪声来自 Noise-92 数据库，含噪语音是在信噪比-5dB 到 15dB 的范围内每间隔 5dB 叠加噪声得到。噪声数据的采样率也为 16KHz，采样比特为 16bit。信噪比(SNR)按下式计算：

$$SNR = 10 \log_{10} \left(\frac{P_s}{P_n} \right) \quad (52)$$

其中 P_s 和 P_n 分别为信号和噪声的线性功率。

软硬件平台

实验程序运行在 Pentium□450 机器上，内存为 128M，选用的操作系统是 Windows 2000。实验使用的抗噪声语音识别系统包括前端增强、特征提取、模型训练、噪声补偿、识别程序和相应的性能评测软件。

识别性能评价标准

对于语音识别系统来说,评价系统性能的主要指标是识别率,也称为识别精度(Accuracy),当然还有其它的一些标准,如识别速度,词汇量大小等。由于我们的实验是噪声环境下的小词汇量连接词语音识别,实验目的是评测各种抗噪声语音识别方法的优劣,因此主要考虑识别率这一项指标。

对于 W_N 个要识别的字,识别系统出现了 W_S 个替代错误, W_D 个删除错误以及 W_I 个插入错误,识别精度(Accuracy)定义为:

$$\%accuracy = [(W_N - W_D - W_S - W_I) / W_N] \times 100\% \quad (53)$$

针对不同噪声:

对特征加权算法和前端 MMSE 增强、Log-Add 模型补偿进行抗噪声性能的比较,特别提出,在我们的特征加权算法中,只对静态特征进行加权;

将特征空间的加权处理分别与信号空间的前端 MMSE 语音增强和模型空间的 Log-Add 模型补偿相融合,分析算法融合后的抗噪声性能;

对本发明提出的 MMSE-FW-LA 方案和 MMSE-LA 方案进行比较。
高斯白噪声(white)

表 1: 高斯白噪声环境下采用不同方法的识别精度

	-5dB	0dB	5dB	10dB	15dB
Baseline	6	8	13	30	65.33
MMSE	14.67	24	54	80.33	91
FW	24.33	46.33	65	76.67	82
LA	22.33	34.33	67.67	84.67	92.67
MMSE-FW	46.33	76	85	91.67	92.67
FW-LA	32	57.33	77.67	87.33	94.67
MMSE-LA	65.33	79.67	89.33	93	93.33
MMSE-FW-LA	48.67	86.67	89.33	94.67	94.67

其中, Baseline 表示未采用任何抗噪声措施的基线系统的识别精度, MMSE、FW 和 LA 分别代表前端 MMSE 增强, 特征加权和 Log-Add 模型补偿。短接符-表示方法之间的融合。

首先我们比较前端 MMSE 增强, 特征加权和 Log-Add 模型补偿方法的抗噪声识别性能, 如图 16 所示:

- 前端 MMSE 增强、特征加权与 Log-Add 模型补偿都改善了噪声环境下的识别性能；
- Log-Add 模型补偿在整个信噪比区间里都优于前端 MMSE 增强；
- 在高背景噪声环境下(SNR<5dB)，特征加权算法是优于 Log-Add 模型补偿的，特别是在信噪比 0dB 时，识别精度提高了 12%；

然后将特征空间的加权处理分别与前端 MMSE 增强和 Log-Add 模型补偿融合，比较它们的识别性能。如图 17 所示：

- 特征加权与前端 MMSE 增强和 Log-Add 模型补偿相融合，相比它们单独处理，都比较明显地提高了识别精度；
- 特征加权与前端 MMSE 增强的融合，在低信噪比时(SNR<15dB)优于和 Log-Add 模型补偿的融合；

比较 MMSE-FW-LA 方案与 MMSE-LA 方案，如图 18 所示：

- MMSE-LA 和 MMSE-FW-LA 方案都显著地提高了噪声环境下的识别精度，在信噪比-5dB 时，MMSE-LA 的识别精度达到了 65.33%，MMSE-FW-LA 更是高达 81%。
- 融合信号、特征和模型三个空间抗噪声语音识别技术的 MMSE-FW-LA 方案优于仅在信号和模型两个空间进行融合的 MMSE-LA 方案，而且信噪比越低，多空间抗噪声技术融合的优势就越明显。如信噪比-5dB 时，MMSE-FW-LA 的识别精度比 MMSE-LA 提高了 15%。

汽车噪声(leopard)

表 2：汽车噪声环境下采用不同方法的识别精度

	-5dB	0dB	5dB	10dB	15dB
Baseline	0.67	17.67	41.33	60.67	80
MMSE	48	73	87	95	96
FW	24	29.33	42	69	84.67
LA	55.33	77.67	93.67	95.33	97.33
MMSE-FW	51.33	74.33	88.67	95.33	96.33
FW-LA	74	86	94	97	97
MMSE-LA	85.67	91.67	95	95.33	96
MMSE-FW-LA	96.67	99	99.67	99.67	99.67

同样，Baseline 表示未采用任何抗噪声措施的基线系统的识别精度，MMSE、FW 和 LA 分别代表前端 MMSE 增强，特征加权和 Log-Add 模型补偿。短接符-表示方法之间的融合。

首先我们比较前端 MMSE 增强，特征加权和 Log-Add 模型补偿方法的抗噪声识别性能，

如图 19 所示:

- 前端 MMSE 增强和 Log-Add 模型补偿都比较明显地提高了识别精度,特别是 Log-Add 模型补偿在整个信噪比区间里都优于前端 MMSE 增强;
- 特征加权与前端 MMSE 增强和 Log-Add 模型补偿相比,识别精度明显下降。主要原因是在特征加权算法没有对无声段语音有效处理,结果导致起伏比较明显的汽车噪声在无声段引入很多插入错误,造成识别率的降低;
- 特征加权与基线系统(Baseline)相比,识别性能还是有所改善。说明在语音的有声段,特征权重的估计和加权处理是有效的。

然后比较特征空间的加权处理分别与前端 MMSE 增强和 Log-Add 模型补偿融合后的抗噪声识别性能。如图 20 所示:

- 与特征空间的加权处理相融合,比较明显地提高了前端 MMSE 增强和 Log-Add 模型补偿的抗噪声识别性能。在信噪比为-5dB 时,识别精度分别提高了 3.33%和 18.67%。
- 在信噪比低于 10dB 时,Log-Add 模型补偿与特征加权的融合效果优于前端 MMSE 增强,这刚好与高斯白噪声环境下的情况不同;

比较 MMSE-FW-LA 方案与 MMSE-LA 方案,如图 21 所示:

- MMSE-LA 和 MMSE-FW-LA 方案在信噪比低于 5dB 时显著地提高了噪声环境下的识别精度,如在信噪比-5dB 时,MMSE-LA 和 MMSE-FW-LA 的识别率都高于 80%;在信噪比高于 5dB 时,也适度地改善了识别器的性能。
- 多融合了特征空间抗噪声语音识别技术的 MMSE-FW-LA 方案优于 MMSE-LA 方案,在 -5dB 到 15dB 范围内,识别精度平均提高将近 2%。

从实验结果可以看出,特征加权算法可以有效的提高低信噪比环境下识别精度,优于前端的 MMSE 增强和 Log-Add 模型补偿;更为重要的是,由于前端语音增强技术、特征加权和模型补偿算法分别针对噪声在信号、特征和模型空间造成的失配进行处理,因此不同方法可以相互融合,整体地提高语音识别系统的抗噪声性能。本发明提出的 MMSE-FW-LA 方案融合了多空间抗噪声识别技术,很大幅度的提高了强背景噪声环境下的识别精度,在 SNR 为 -5dB 的高斯白噪声和汽车噪声环境下,识别精度都达到了 80%,而且从算法复杂度来看,MMSE-FW-LA 方案的前端增强和特征权重估计相互融合,选用了计算量较低的 MMSE 估计方法,模型补偿不需要对噪声模型进行离线估计,这些都有利于此方案的实时处理。因此,本发明提出的 MMSE-FW-LA 方案具有很强的实用性。

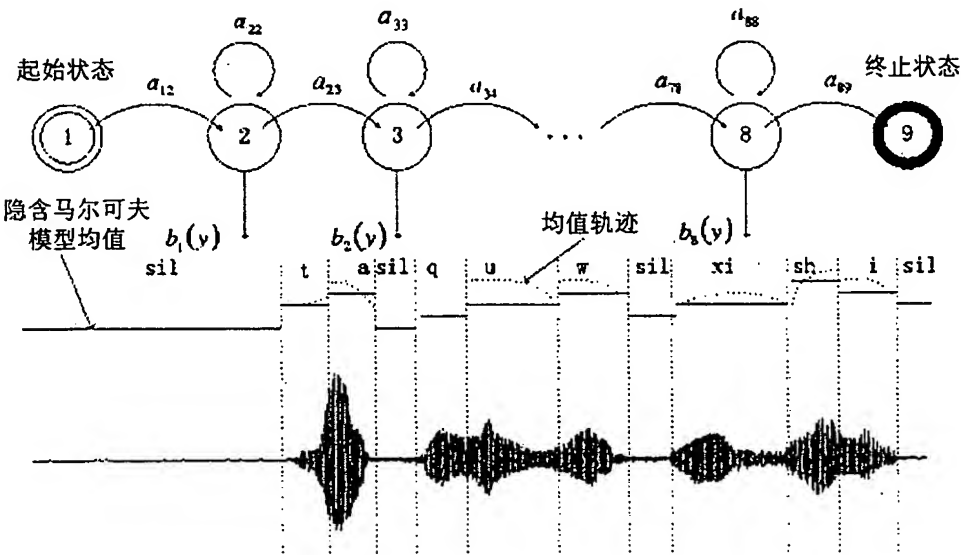


图 1

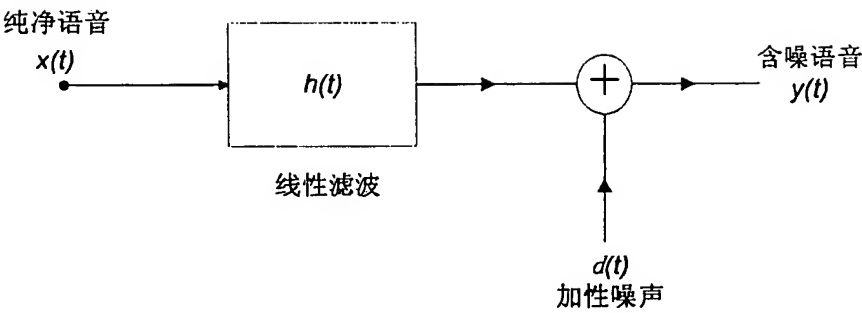


图 2

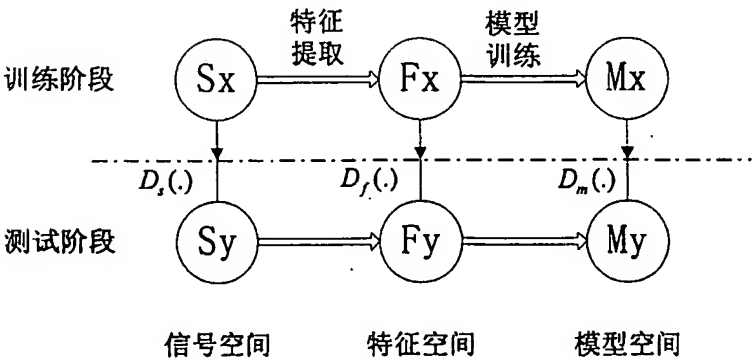


图 3

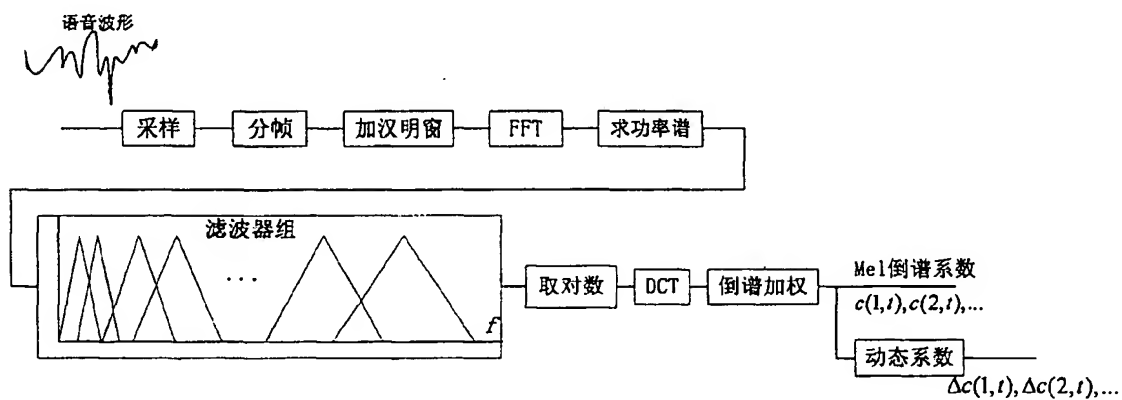


图 4

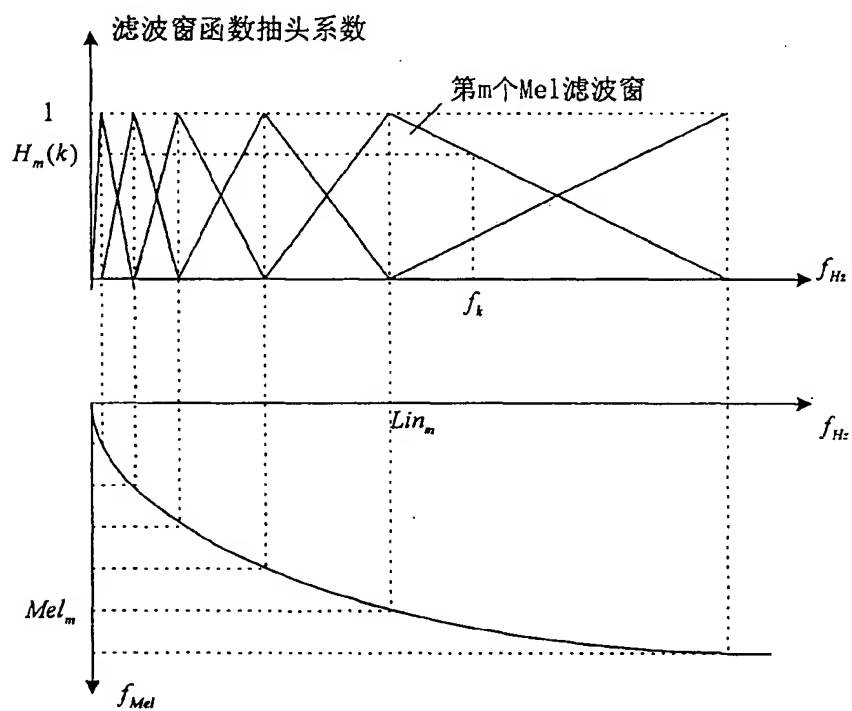


图 5

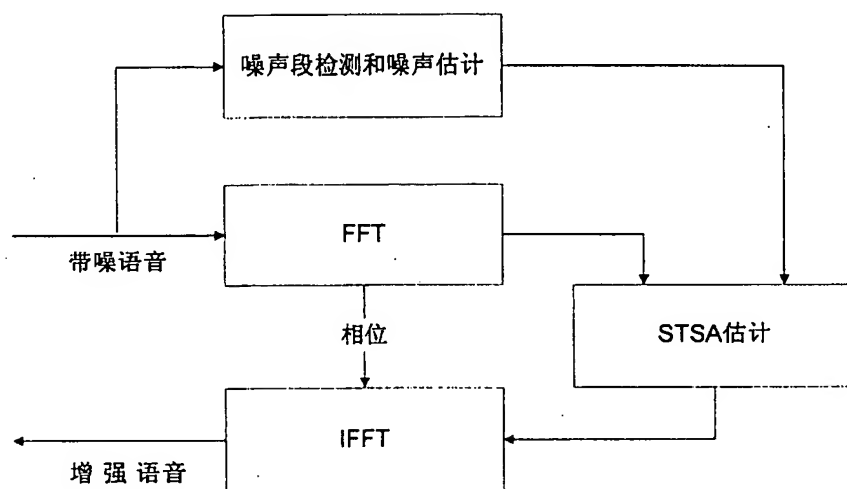
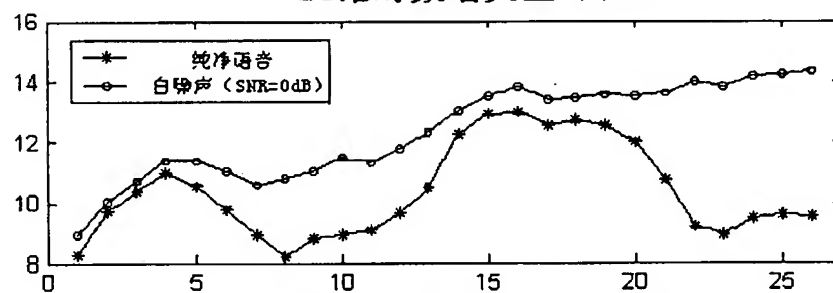


图 6

26维对数谱矢量 图(a)



26维对数谱矢量权重 图(b)

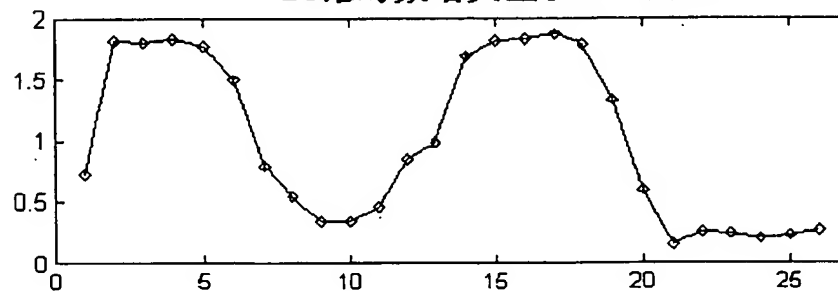


图 7

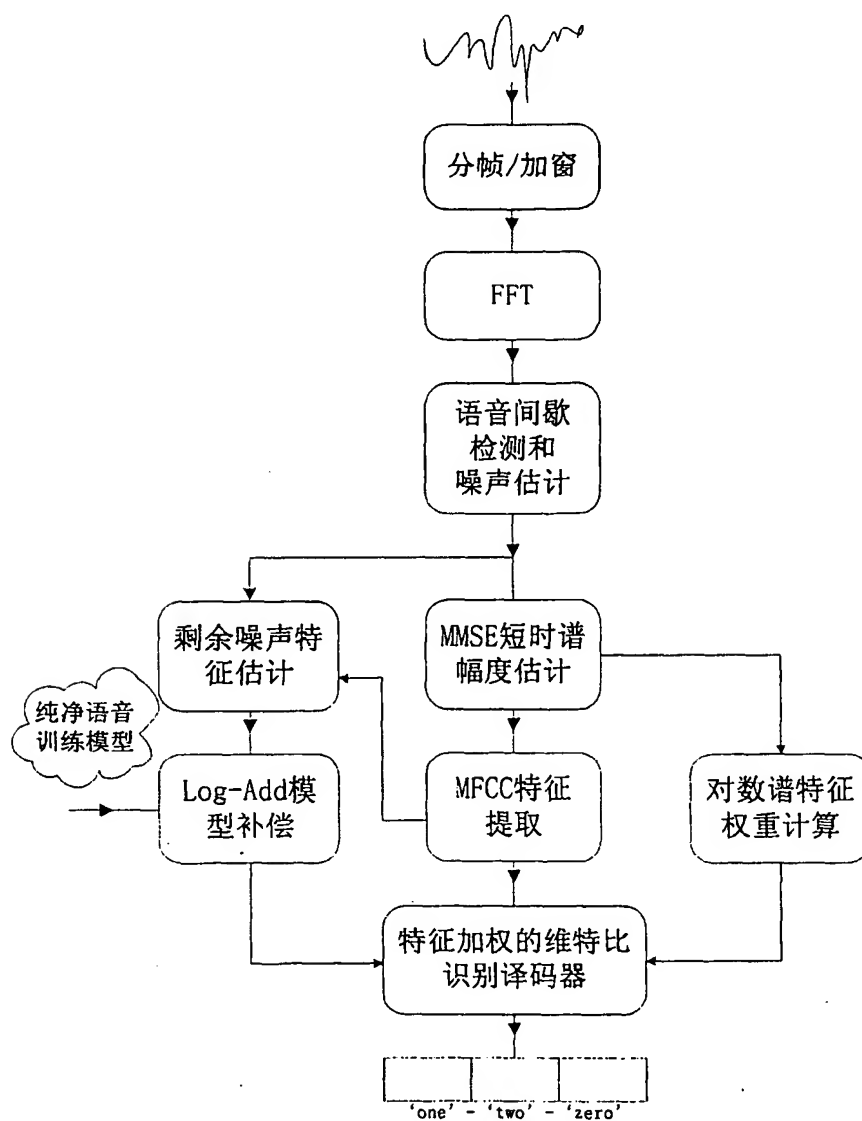


图 8

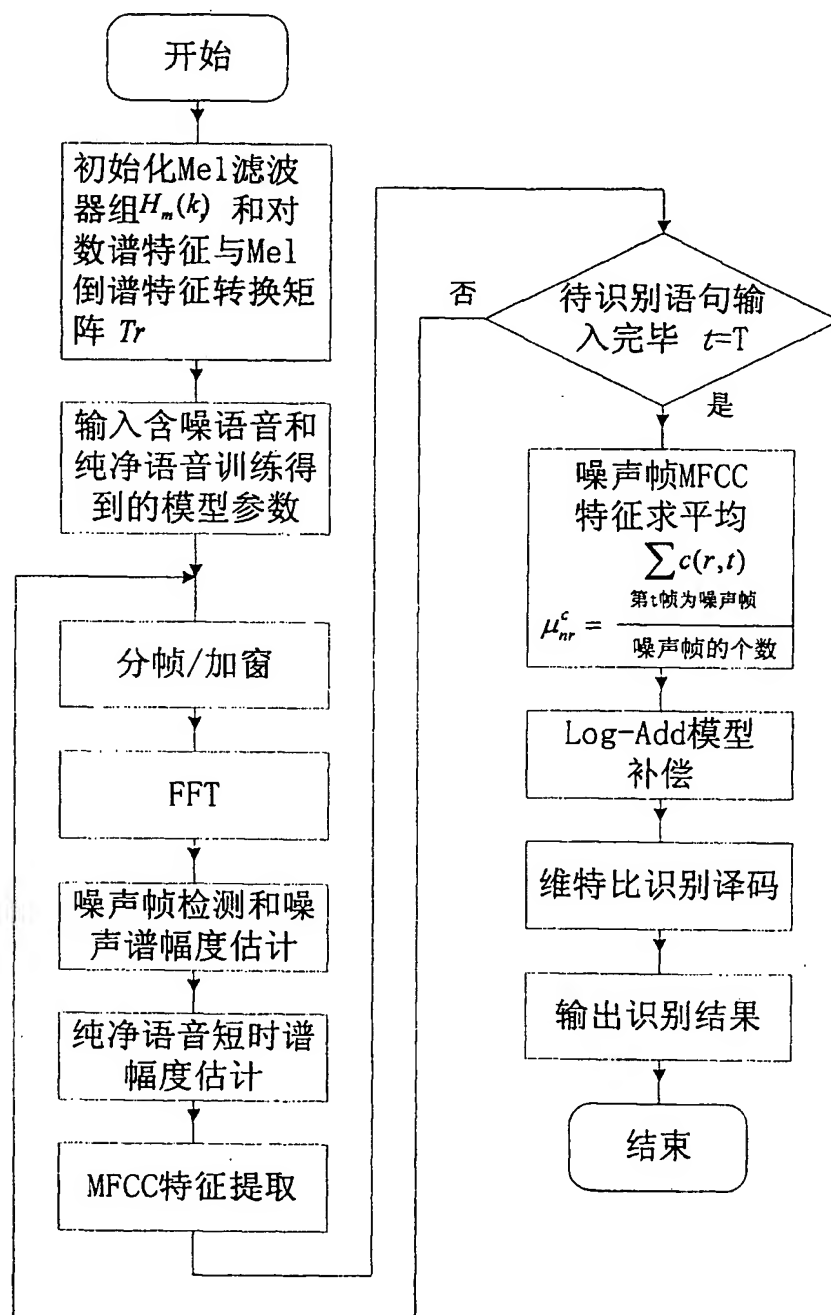


图 9

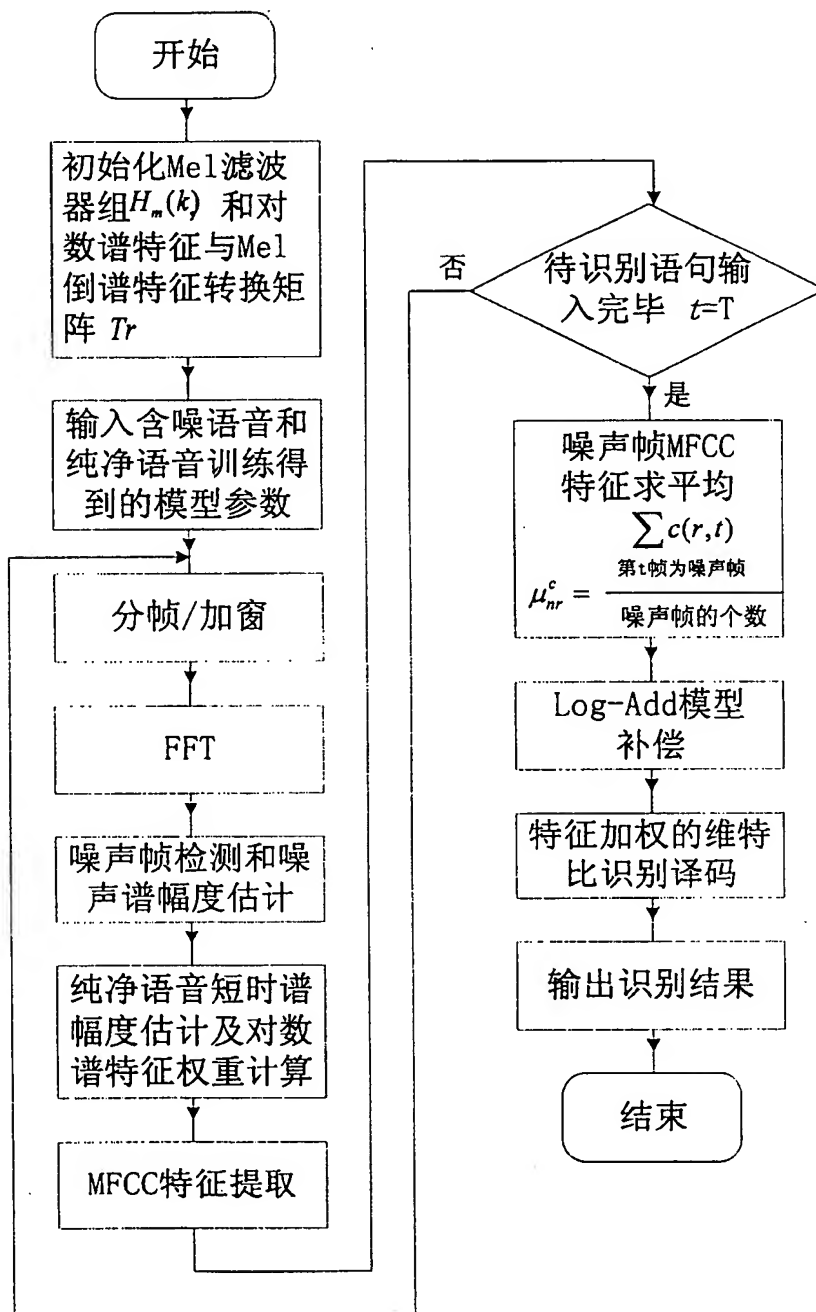


图 10

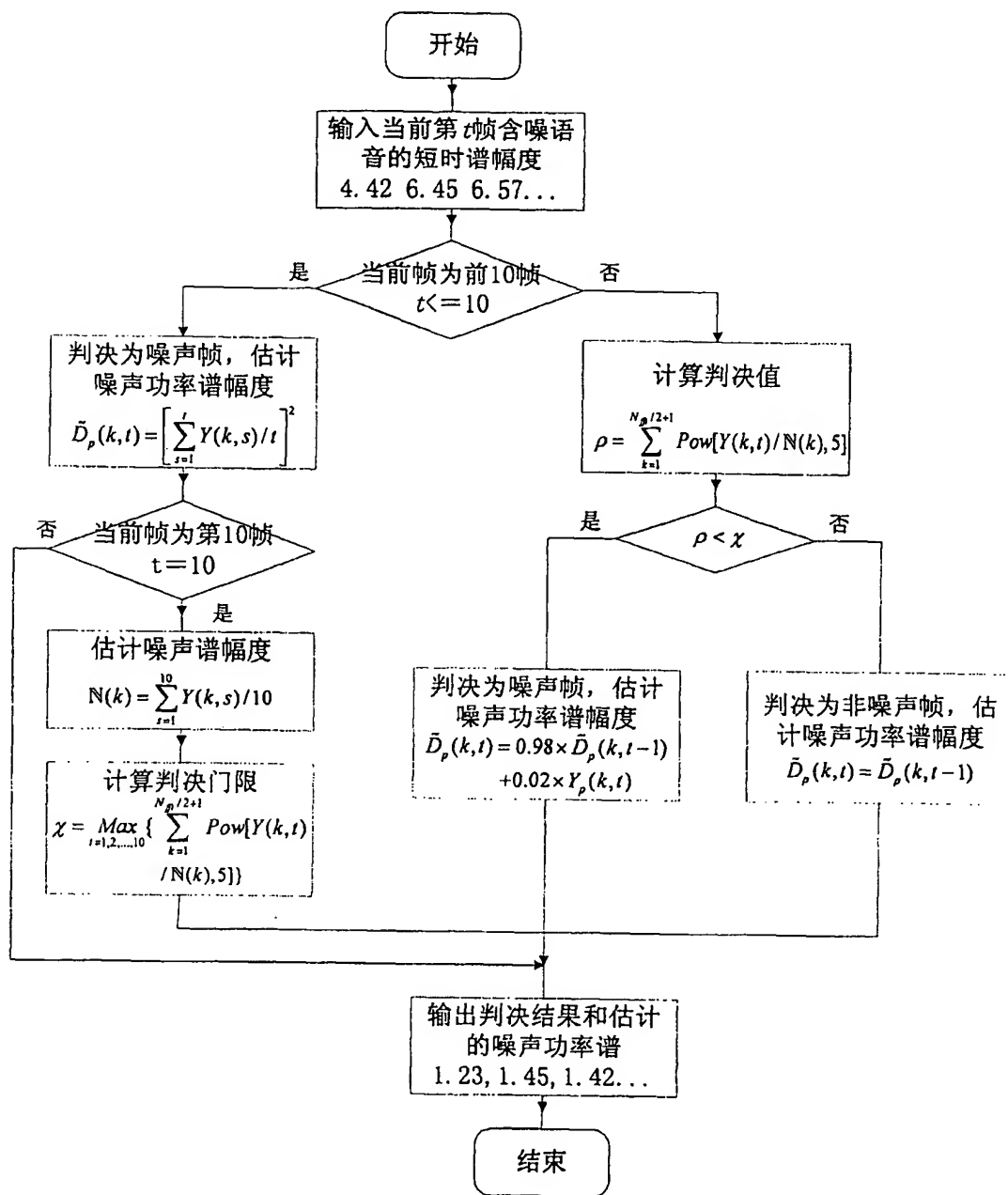


图 11

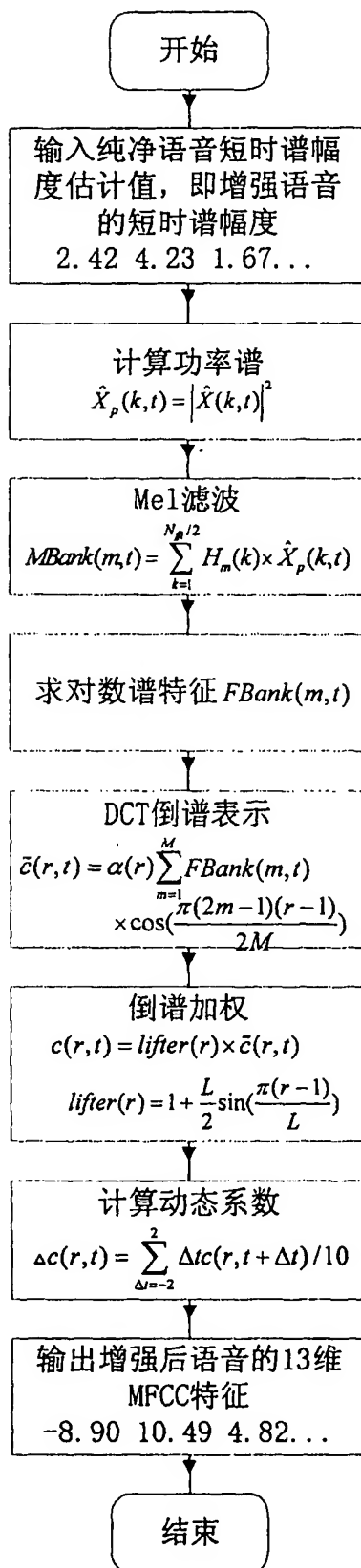


图 13

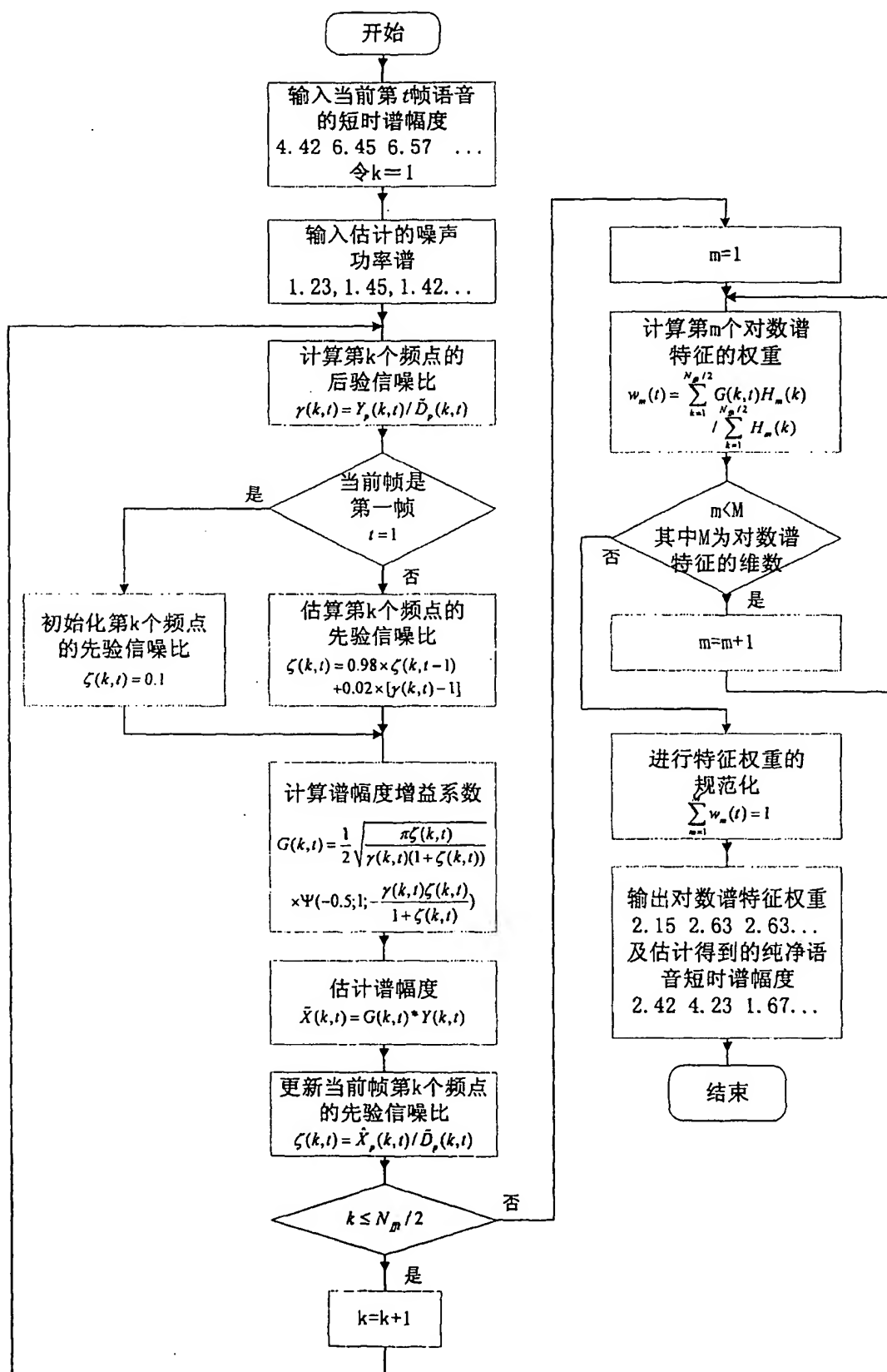


图 12

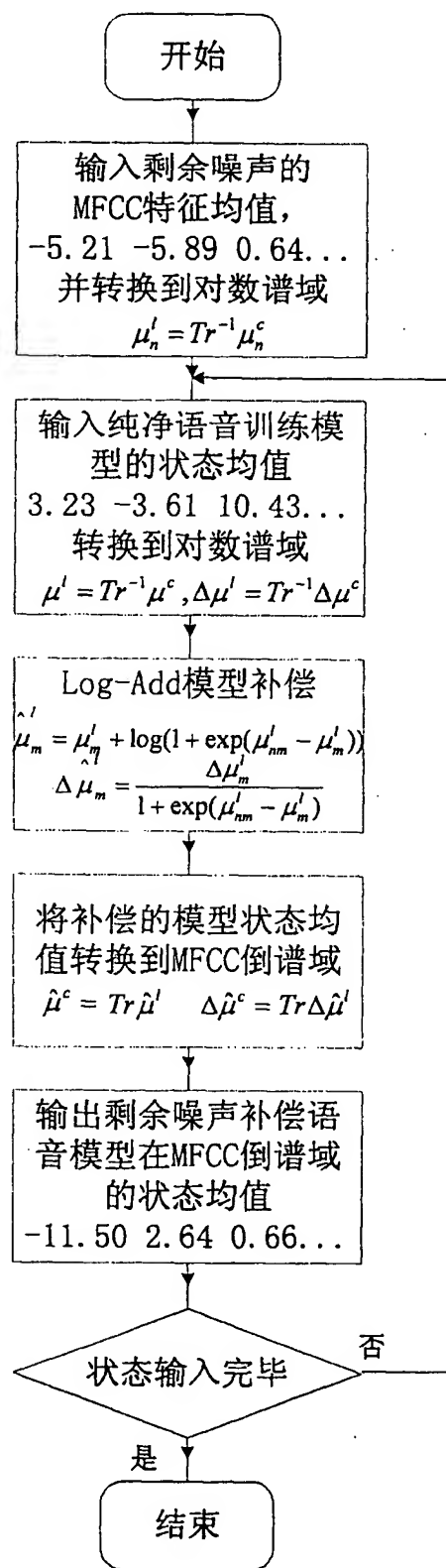


图 14

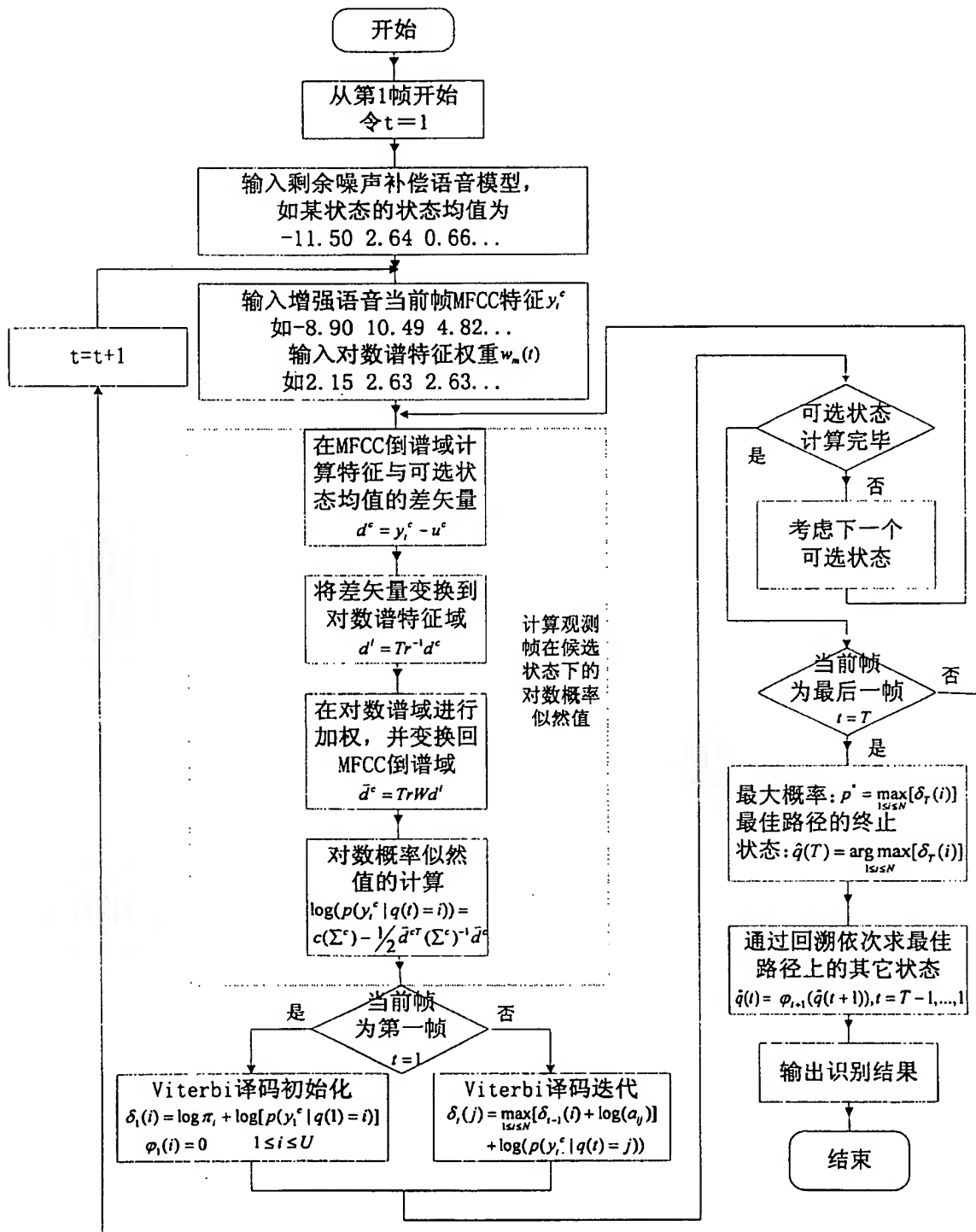


图 15

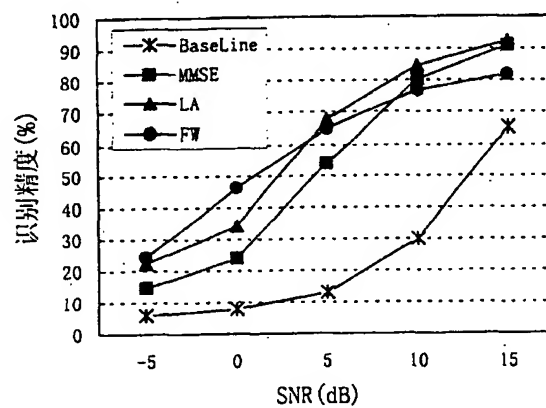


图 16

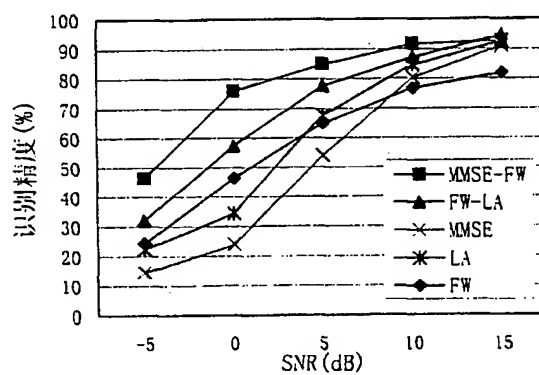


图 17

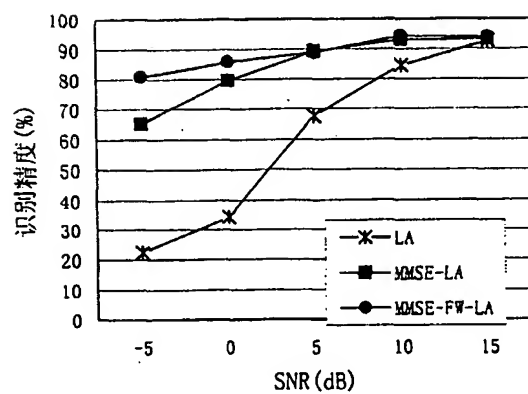


图 18

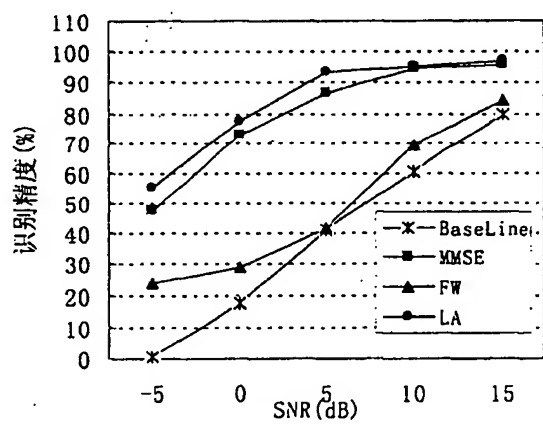


图 19

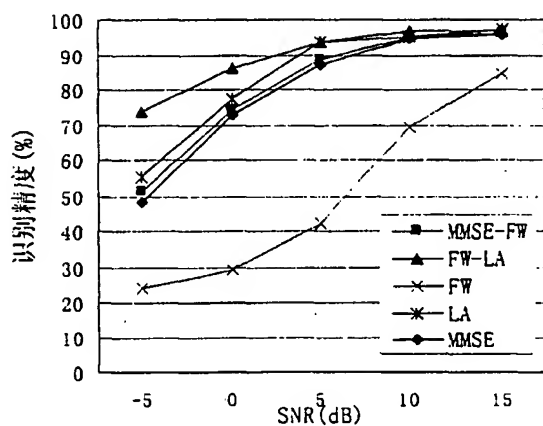


图 20

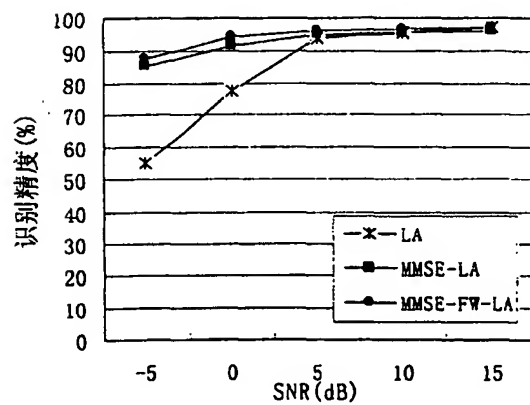


图 21